High-speed Sampler for CMOS LIDAR

Emil Ulvestad

Thesis submitted for the degree of
Master in Electronics and Computer Science, Microelectronics
60 credits

Department of Physics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Autumn 2018

High-speed Sampler for CMOS LIDAR

Emil Ulvestad

High-speed Sampler for CMOS LIDAR

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Preface

This Master's thesis was submitted for the degree of M. Sc. at the University of Oslo, dept. of Informatics. The work has been carried out in the period between spring 2017 and fall 2018 at the NANO group under the supervision of Professor Tor Sverre Lande, Ph. D. Kristian G. Kjelgård and Professor Dag T. Wisland.

Oslo, August 15, 2018

Emil Ulvestad

Emil Ulvestad

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my supervisors, Tor Sverre Lande, Kristian G. Kjelgård and Dag T. Wisland. Your continuous guidance and scientific insights during our countless discussions have been invaluable. You have shown great enthusiasm and patience in my work, which has allowed me to develop both my knowledge and confidence.

I would also like to thank Øystein Bjørndal for your undisputed expertise in Cadence, Mathias Tømmer for your assistance in making the Beaglebone Black behave and Olav Kyrvestad for your practical assistance during tape-out and PCB production. A special thanks go to my partner in crime, Tohid Moradi Khanshan, for sharing the chip area with me and keeping me company during the long days in the lab.

I am also very grateful for the support that my friends and family have shown. Last but not least, I would like to thank my girlfriend, Vibeke, for always having faith in me and supporting and fuelling my ambitions without questions.

E.U.

Abstract

In this thesis, a high-speed sampling system intended for use in a depth selective spectroscopic LIDAR for transcutaneous blood assessment is proposed. The system is fundamentally based on a continuous-time sampling solution, enabling sampling rates close to $100GS/s$, close to two orders of magnitude faster than standard clocked systems. Oversampling makes the system capable of capturing inherently weak and noisy high-frequency signals. Based on experimental verification in 90nm CMOS technology, the opportunities and limitations of continuous-time sampling systems are analyzed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the increasing portability and computational power of modern microelectronics, the interest towards development of new non-invasive biomarker sensing technology has increased significantly. Technology advances have enabled many wearable solutions providing long-term monitoring of cardiac, respiratory and metabolic diseases, like for example diabetes. However, most chemical monitoring techniques, e.g., blood glucose analysis for diabetes patients, remain subcutaneous[1]. In order to achieve entirely non-invasive detailed chemical assessments of biomarkers, *transcutaneous* sensing[2] is being explored as a potential solution. Pulse oximetry, in which near-infrared (NIR) light is passed through some thin part of the body and sensed by a photosensor on the opposite side, is a well established transmissive transcutaneous sensing technique used for measuring blood oxygen levels (hemoglobin saturation). However, oximeters do not provide sufficient accuracy for more specific monitoring of substances in the blood like glucose or carbon dioxide levels.

Spectroscopy is a well established scientific measurement technique for characterizing matter by analyzing its interaction with electromagnetic radiation (EMR). As EMR interacts with matter, it is either transmitted, absorbed, reflected, scattered or undergo photoluminescence[3]. Each atom and molecule render a unique response, and the resulting spectrum can be used to recognize the biochemical composition of the sample. As an example: Oxyhemoglobin (HbO_2), which is a type of hemoglobin found in blood, show high levels of absorption of light between $300nm$ and $600nm$. For the sake of simplicity, assume that no other substances in the sample share this characteristic. By emitting EMR in this spectrum, and studying the results, one can determine the levels of HbO_2 in the sample by the degree of attenuation. Near-infrared (NIR) defines light ranging from $650nm$ to $1350nm$ in wavelength, and is where light has its highest depth of penetration in body tissue. However, with a predominant absorption of hemoglobin at wavelengths $< 600nm$ and water in the wavelength range $> 1\mu m$, plus the fact that silicon substrate photosensors only show adequate sensitivity up to $\sim 1\mu m$ biochemical characterization beyond a millimeter of tissue becomes impractical when outside the wavelength window of $\sim 600 - 1\mu m$ [1]. This window is often referred to as the *therapeutic window*. Although NIR light has high penetration in tissue, spectroscopy does not automatically become particularly accurate. Within the theurapeutic window the dominant light-tissue interaction is scattering, which causes rapid diffusion. The scattering increases the distance traveled by photons, which in turn leads to a higher probability of photon absorption. With the high degree of absorption in the body tissue, obtaining high *sensitivity*, i.e., signal-to-noise ratio (SNR), becomes a significant challenge in transcutaneous light sensing. Complex interactions between light and tissue further complicate the task of assigning specific properties to samples. As an example; when attempting blood glucose measurements, the weak NIR spectra from glucose is severely obscured behind the stronger overlapping spectra from water, hemoglobin, proteins, and fats in the human body. Estimating the glucose concentration from the scattered light, therefore, becomes challenging. With such broad and often overlapping bands, achieving high sensitivity in spectroscopy is not trivial.

Some advanced spectroscopic systems using optical techniques to provide depth profiling in semi-transparent materials have already been developed, as shown by Freebody et al. [2]. However, these systems are large, expensive and consist of complex optical solutions unsuitable for microelectronics. In addressing the challenges regarding selectivity, solutions in time-of-flight (ToF) measurements hold promising prospects when combined with microelectronics. With directed measurements of scattering at specific depths, the selectivity

[1] Subcutaneous sensing: Sampling from tissue located under the skin.

[2] Transcutaneous sensing: Sensing measured across the depth of the skin.

[3] Photoluminescence designates a number of effects, including fluorescence, phosphorescence, and Raman scattering.

may be significantly improved. A well-known ToF instrument is the radar. A simple radar solution is pulsed radar, which transmits a short radio pulse, at which point a high-speed a sampler is used to measure the ToF from the backscattered signal. This principle is also applied using light pulses for long-range sensing, conventionally named Light Detection and Ranging (LIDAR). Seeking existing solutions, we find many such light-based time-of-flight systems already integrated into CMOS technology, like the laser autofocus, found most high-end mobile cameras. However, such systems are either too large [3] or lack sufficient depth resolution [4].

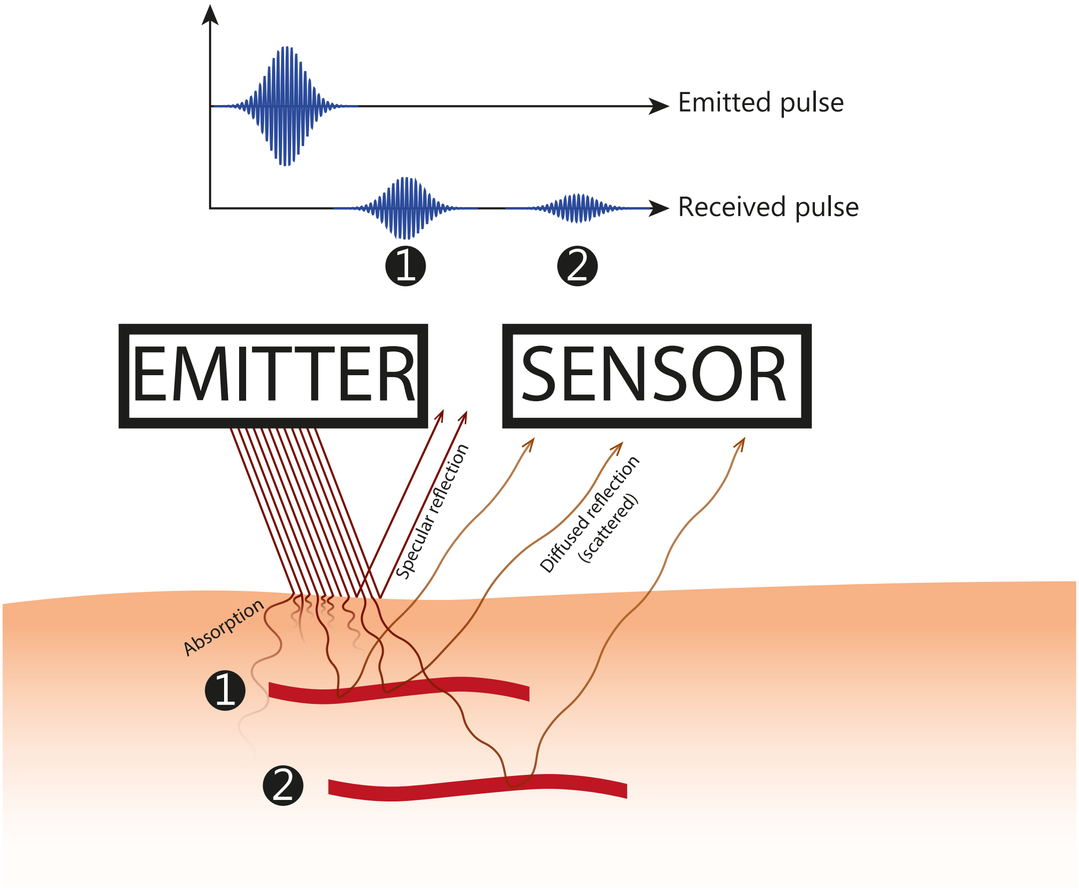


Figure 1.1: ToF measurements will provide depth profiling which might increase selectivity in transcutaneous sensing.

The phrase *"A chain is only as strong as its weakest link"* is appropriate in describing how clock synchronized systems must limit their speed. The slowest logic stages in such a system will dictate the maximum clock frequency to avoid catastrophic race conditions which could lead to incorrect data or glitches. There will be uncertainty regarding arrival times of clocked signals, so special consideration must be made to meet the timing requirements with an acceptable margin. This limitation in speed restricts the depth selectivity when processing light signals. The diameter of a typical blood vessel in the underarm is in the order of $1mm$. Sensing properties of the blood inside such a vessel will have to approach the same order of resolution. With light propagating through the body in the order of $10^8 m/s$, we find that a $1mm$ resolution will require a sampling rate of approximately $100GS/s$ ($\frac{10^{-3}m}{10^8 m/s}$). To this time, such high frequencies are considered unreachable in conventional microelectronics as most clock synchronized commercial systems have a peak clock frequency of no more than $5GHz$. However, the continued development of microelectronics has brought with it significant improvements with regards to transistor sizes which has resulted in lower capacitances and faster digital gates. A typical gate-delay in nanometer technology is around $10ps$ and is still becoming faster with further advancements. These progressions have led to a new 1-bit signal processing domain; Continuous-Time Binary-Value (CTBV), which as a concept is based on the idea of substituting the clocked synchronization with delay-based timing. By removing their dependency on the clock, systems can operate in continuous time, with speeds limited by the inherent gate delay of the given technology.

The primary objective of this thesis is to explore how a high-speed signal acquisition system capable of reconstructing inherently weak and noisy high-frequency signals at an equivalent sampling rate approaching $100GS/s$ can be realized in standard CMOS technology. The architecture will be fundamentally based on CTBV signal processing. We will explore how CTBV allows us to regain the speed scaling provided by the technology, by utilizing the inherent gate-delay to design a temporally sequenced sampling system of interleaved delay lines. However, with great speed comes great responsibility. The stochastic process variations and temperature dependencies in CMOS introduce behavioral differences between devices which must be accounted for both before and during operation. We will rely on careful modeling and also explore back-gate tuning as a method of compensation during operation. However, the process variations are expected to become a limiting factor of the timing precision of signals and will add restrictions on both speed and depth of the system. In this thesis, we will be pushing the boundaries of the technology, and we hope through empirical, experimental verifications to reach definite conclusions regarding the potentials and limitations of continuous-time sampling and make way for future work.

Chapter 2

Background

In this chapter we will explore the theoretical foundation for realizing the sampling system, before presenting the proposed system architecture. Following comes a discussion on the challenges this introduces.

2.1 1-bit signal processing

As the supply voltage and headroom of fine-pitch technologies decrease, implementing high-speed analog signal processing is becoming a significant challenge. The low power supply voltage prohibits any high performance integrated analog processing. 1-bit signal processing, on the other hand, has become a strong candidate, since it takes advantage of the many inherent properties of fine-pitch digital technologies; such as high speeds, simple design, low cost and a high degree of flexibility. Moving away from analog processing to multi-bit digital processing presents many potential benefits in itself, such as:

- Implementations of all-in-one packages on small Silicon die (SoC).
- Allows for more advanced and flexible systems.
- Less deterioration during transmission and write/read. Data can be stored in memory and transferred without loss.
- More immune to noise, process- and environmental variations.
- There is also a potentially decreased power consumption since digital electronics have a much lower static power consumption compared to analog electronics.
- Operations can be parallelized.

Venturing from multi-bit to 1-bit signal processing there are additional benefits to be gained:

- 1-bit signal processing enables simpler, smaller designs, without the need for parallel data buses. With serial buses, timing skew in buses is no longer of any concern (allows for even higher speeds).
- Compared to multi-bit, 1-bit processing have potential for even lower power consumption.
- 1-bit AD/DA conversion will be much simpler and have a higher linearity than multi-bit converters. Multi-bit converters have many threshold levels and great care have to be made to ensure linear response. A 1-bit converter is either 0 or 1, so it can't have unevenly spaced levels.
- As highlighted in the next section, CTBV systems allow for high temporal resolution with the removal of the clock.

1-bit signal processing does, however, come with some limitations and potential drawbacks. The following points are considerations made from the perspective of our sampling system. Firstly, even though the system is not dependent on clocked synchronization, a clock will be required to trigger the sampling. Secondly, some form of oversampling or integrating action will be required to compensate for the inherently low dynamic range (DNR) and SNR of a 1-bit quantized system. As we will see in section 2.3 we achieve this through signal averaging, in which consecutive samples can be used to increase DNR and SNR. However, this oversampling will come at the cost of increased processing time and can be power consumption. Thirdly, several "analog" operations, such as filtering and clipping are more challenging to achieve in 1-bit systems [5].

2.2 CTBV Signal Processing

As mentioned in chapter 1 the sampler in this thesis is intended to perform readout in a pulsed LIDAR system, where it samples reflected EM pulses. A significant effort must, therefore, be placed in generating very high sampling rates to acquire the high-resolution sensing necessary to obtain a high degree of depth selectivity. The term *Continuous-Time Binary Value*, first coined by Hjortland and Lande [6], represents a unique hybrid signal-processing domain in which the signal is represented in binary, while time is kept continuous through the removal of a clock. By being clockless CTBV can offer the speed of conventional analog signal processing, while also supporting some of the advanced processing capabilities seen in digital signal processing (DSP) due to its binary value discretization. In the CTBV domain, a signal is expressed in a binary manner and is continuous in time. In contrast to clock synchronized digital logic, where transitions are dependent on the clock cycle, CTBV signals can transition at any point in time. This means that the signals are represented primarily by their transitions and pulse widths (similar to PWM[1]), instead of the conventional representation of amplitude level at discrete steps.

		Time	
		Discrete	*Continuous*
Value	*Binary*	Digital	CTBV
	Continuous	Sampled analog	Analog

Table 2.1: Signal processing domains

In the CTBV domain, signals can be processed with a set of simple digital operations, which will be represented in short below.

1. *Delay and pulse shaping:* Delay elements can be constructed as two cascaded inverters. As the signal passes through a delay element, its transitions will be delayed by a predefined amount. If designed symmetrically the inverters will not significantly distort the pulse widths of the signal (some phase noise). However, asymmetric sizing can be applied intentionally to create a delay element which delays the rising or the falling edges of the signal significantly more than its counterpart. This allows us to adjust the widths of pulses easily.

   By cascading multiple delay elements, we create a *delay line*, which provides a sort of history of the input signal within a small time frame. The delay line operation is central to the design of the sampler and is addressed more in-depth in section 2.5.

2. *Sampling:* Simple 1-bit digital samplers can be used to sample CTBV signals at high speed. In this thesis, D flip-flops are used, in conjunction with delay lines, to perform the high-speed sampling, in which the D flip-flops' clock inputs are tapped from the different stages of the delay lines.

3. *Combinational logic:* Due to the binary quantization of CTBV, combinational logic operations (AND,OR,XOR, etc.) become trivial to implement. Such logic operations can, for example, be combined with delay elements for pattern detection, pulse width modulation, etc.

The simple architecture of CTBV signal processing circuits provides opportunities for designing highly area efficient systems. Also, with the exclusion of high power analog system components, and the fact that digital CMOS circuits primarily draw power during switching (dynamic power), CTBV presents unique opportunities for specialized low-power designs (which has become increasingly relevant in the recent decade). In the absence of the clock, many CTBV operations become limited mainly by the gate delays in the technology used - which in modern CMOS technologies can be as low as a few tens of picoseconds. A simple, yet clever technique to further increase speeds beyond the gate delays is easy to implement. The relative difference in delays

[1] In pulse-width modulation the amplitude(s) of a signal (usually analog) is encoded into the pulse width(s) of another signal (usually digital)

can be utilized to perform operations which will have a time resolution much smaller than the individual gate delays. This technique is of fundamental relevance in this sampler, but as we will address in subsection 3.2.2, manifests a significant design challenge due to device mismatch. In addition to increased speeds, the absence of the clock can further reduce power consumption and frees up a significant portion of the footprint which is usually intended for clock distribution. Although, it should be mentioned that clocked logic is often necessary for further processing following high-speed CTBV operations. Luckily this can be done off-chip or in isolated parts of the system in a sort of hybrid design.

Quantization in digital signal processing involves the mapping of the continuous (time and value) input signal to a *limited* number of levels (bits) at discrete time intervals. As a rule of thumb, when a continuous signal is to be digitized, the finer the quantization, the more accurate the reconstructed signal will be. However, there is bound to be some error. This error, referred to as the *quantization error (QE)*, describes the difference between the input value and its quantized value originating from the limited resolution of both time and value. In a CTBV system quantization error is eliminated as a problem as long as the signal remains continuous. In our system, we are quantizing and sampling in two separate operations. The quantizer used in this sampling system outputs a continuous-time binary signal that transitions precisely as the input signal crossed the given threshold level. We see that by not discretizing time (or value), no quantization error is introduced. However, upon sampling the 1-bit quantized signal, time is discretized, and the quantization error is introduced into the system.

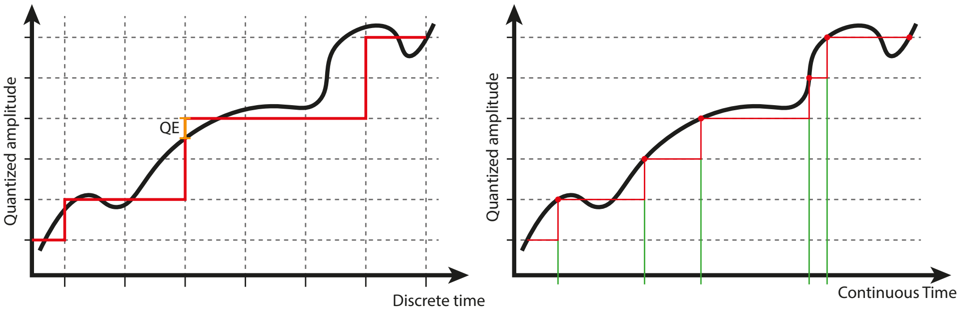


Figure 2.1: QE is eliminated by removing discretization in either ampl. or time.

Many of the aforementioned advantages of CTBV signal processing also highlights some of its shortcomings. It is its simple logic and clockless design that facilitates the superior speeds, but it also limits the coding scheme of signals to just one bit, restricting available on-chip operations. Also, CTBV signal processing places a significantly harsher demand on the designer. For example, with poor design, it is conceivable that short pulses can completely disappear while propagating through a CTBV system, e.g., a delay line. If the inverters in said delay line are asymmetrically designed or have significant device mismatch, one transition will propagate faster through a gate than the other, effectively changing the pulse width. In a deep delay line, this possibly unintentional pulse shaping can eventually lead to pulses shrinking or expanding and eventually disappearing as one transition catches up to the other. At these high speeds, the performance also becomes more vulnerable to the inherent process variations in CMOS, i.e., *device mismatch*, which is the subject of section 2.6. By relaying information within the signal transitions and pulse widths, CTBV effectively becomes all the more vulnerable to this naturally occurring variation. We will be pushing the technology (TSMC 90nm low-power CMOS) towards its lower size limits to achieve the needed gate delays. A direct consequence of this will be a higher presence of mismatch. Considerable attention must be on reducing or compensating for the resulting effects of mismatch if we are to create a stable and robust sampling system.

2.3 Signal averaging

When there is a significant occurrence of noise within the frequency band of the signal, it will obscure the signal. With significant overlap between the signal and noise spectra, conventional noise reduction techniques, such as filtering, will fail since the filtering will also attenuate frequency components present in the signal, distorting it. Also, as mentioned in chapter 1, the backscattered signal will be very weak due to the high level of absorption in the body, leaving very little to be sensed by the photosensor. In other words, we have an inherently weak reflected signal buried in noise which is to be represented with only a binary value quantization. Without compensation, the resulting SNR would be unacceptably low ($< 0dB$). Fortunately, the high speeds and binary coding scheme makes CTBV signal processing an ideal candidate for oversampling techniques such as signal averaging. Signal averaging (or processing gain), which is applied in the time domain, will improve the SNR and resolution of an A/D system at the cost of increased processing time and reduced throughput. By averaging over the sum of replicate measurements we can achieve an SNR significantly higher than the low native SNR of 1-bit signals. This is because the noise will average out with iterations, while the signal waveforms will sum together. There are however some requirements for this method to work. Signal averaging relies on repeated measurements, so the signal must be coherent (but not necessarily periodic). Also, for signal averaging to be effective, the signal must contain uniformly distributed noise (white) with a sufficient amplitude to cause random changes comparable to at least one LSB between samples. It is the random nature of white noise that makes signal averaging effective. By aligning each measurement with the previous iteration(s), the waveforms are summed. For this to work the temporal position of each signal waveform must be accurately known. The systematic signal components, S, in each repetition are combined so that after, e.g., m repetitions, the signal amplitude is m times larger than for a single measurement (mS). The noise, however, assuming that it is random, with $\mu = 0$ and an average RMS value σ_n, will after m repetitions be the square root of the sum squared ($\sqrt{m\sigma^2} = \sqrt{m}\sigma$). Thus, the SNR is improved by $\sqrt{m}$. This is confirmed mathematically below:

Assume an input signal $f(t)$ consisting of both the coherent signal $S(t)$ and the noise $N(t)$:

$$f(t) = S(t) + N(t) \tag{2.1}$$

$f(t)$ is sampled at a time interval of τ. The value of a sample in time ($i = 1,2,..,n$) is then the sum of the signal and noise at that sample point.

$$f(\tau_i) = S(\tau_i) + N(\tau_i) \tag{2.2}$$

For each repetition the data is accumulated in memory. So after m repetitions we have:

$$\sum_{k=1}^{m} f(\tau_i) = \sum_{k=1}^{m} S(\tau_i) + \sum_{k=1}^{m} N(\tau_i) \qquad for\ i = 1,2,...,n \tag{2.3}$$

With perfect alignment of repeated measurements, the signal component of each sample point is the same at each repetition, assuming a stable signal.

$$\sum_{k=1}^{m} S(\tau_i) = mS(\tau_i) \tag{2.4}$$

Under the assumption that the noise is uncorrelated and random with a mean of zero: after multiple repeated measurements $N(\tau_i)$ has an RMS value of σ_n.

$$\sum_{k=1}^{m} N(\tau_i) = \sqrt{m\sigma_n^2} = \sqrt{m}\sigma_n \tag{2.5}$$

Eqs. (2.4) and (2.5) gives the new SNR after m measurements as:

$$SNR_m = \frac{S(\tau_i)}{N(\tau_i)} = \frac{mS(\tau_i))}{\sqrt{m}\sigma_n} = \sqrt{m} \times SNR \tag{2.6}$$

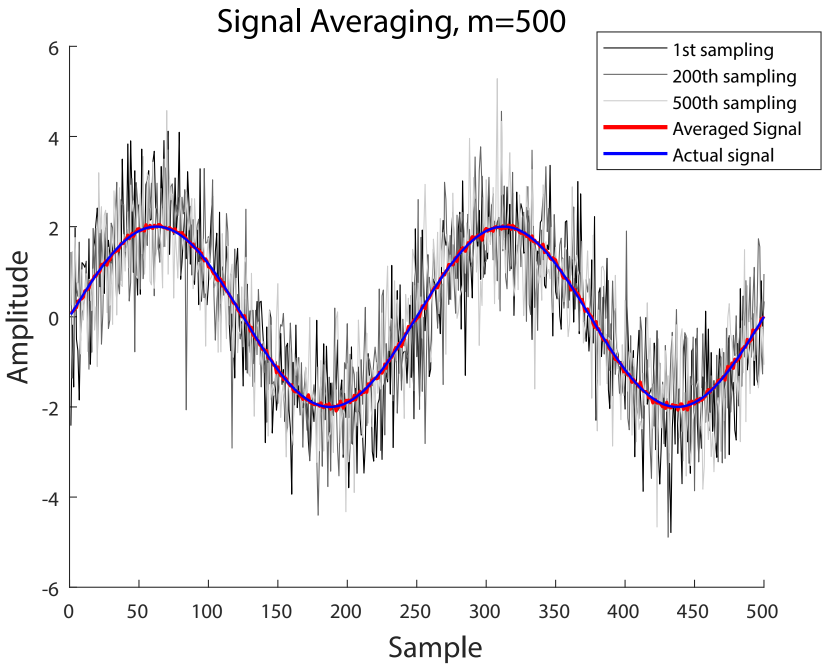


Figure 2.2: Signal Averaging principle illustrated. Severely noisy signal recovered with with signal averaging over 500 measurements

In practice, signal averaging allows us to obtain higher-resolution A/D conversion with increased SNR through oversampling instead of increased complexity. For instance, to implement a 16-bit A/D converter, it can be sufficient to use a 12-bit converter that runs at 256 times the target sampling rate. By combining 256 consecutive 12-bit samples (within the required sampling period), the coherent signal increases by a factor of 256, while the noise by a factor of $\sqrt{256}$, so the SNR changes by a factor of 16 (= 2^4), effectively adding 4 bits to the A/D converter, producing a single sample with 16-bit resolution. However, signal averaging is only effective when the signal contains white uncorrelated noise with an amplitude significant enough to be detected by the A/D converter and causing random quantization level changes. If the A/D converter sees a stable input signal, all samples and the resulting average would be the same, rendering the averaging ineffective [7].

2.4 Swept Threshold Sampling

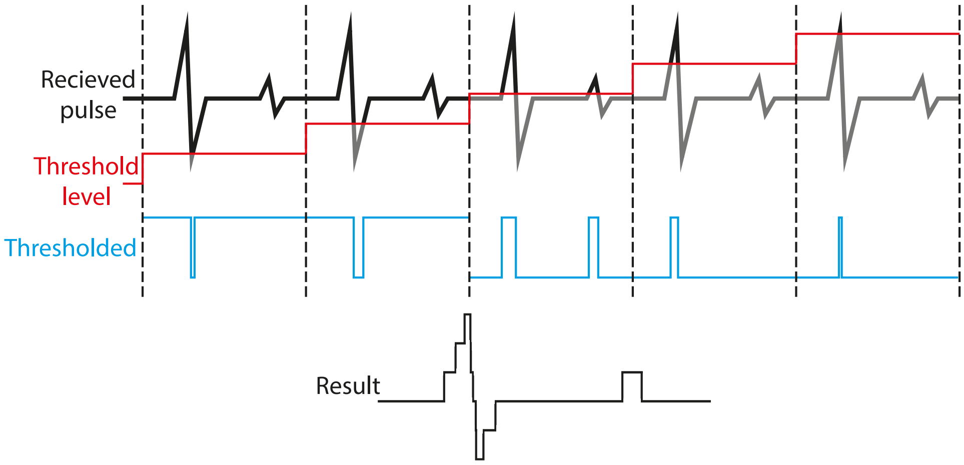


Figure 2.3: Swept threshold principle

The required dynamic range can change significantly depending on the characteristics of the backscattered signal. With a weak and noisy backscatter in the pulsed LIDAR system, swept threshold becomes a viable sampling technique able to recover the pulse information efficiently. In swept threshold sampling a 1-bit quantizer compares the received signal with a given threshold level and outputs a digital, yet still continuous signal. This digital signal is then sampled by a series of 1-bit samplers in quick succession. Since the signal remains continuous after quantization, each sampler will capture the value at the exact time it is triggered. Following the sampling, accumulators connected to each sampler are incremented if the individual sample was "1". This procedure is repeated in a threshold level sweep across (only) the full range of the input signal, resulting in a "sample and accumulate" operation which provides the averaging action needed to reconstruct the signal with a comfortable degree of SNR improvement. As highlighted in the previous section, the signal must be coherent. In other words, the signal (without noise) must be the same between measurements. Considering that the signal to be measured is the reflected pulse, this means that we require a stationary scene of a slowly changing process during the integration period. Furthermore, if swept threshold is to be effective in reconstructing weak signals, adequate threshold level control is required[2]. With swept threshold, the input range and threshold steps can be adjusted to better accommodate to the input signal characteristics. The input range can be traded in favor of faster sweep times, while processing gain can be adjusted with the number of steps [8]. This way the dynamic range of the sampling system will better relate to the input signal.

The theoretical limit of the SNR of an A/D converter (ADC) with no oversampling is:

$$SNR_{dB} = (6.02 \cdot ENOB) + 1.76 \tag{2.7}$$

where $ENOB$ is the effective number of bits. With a small signal amplitude expected from the LIDAR, ENOB will only be equal to the a small portion of the full range of the ADC; the dynamic range of the input signal does not match the rail-to-rail range of the ADC, hence it provides a poor initial resolution and low SNR. A substantial averaging effort will be required in order to acquire a decent SNR, and with a large number of unused quantization levels, the ADC will consume high amounts of power needlessly during this averaging. Swept threshold provides a more targeted A/D conversion. Considering that with weak signals with low SNR signal oversampling is unavoidable, swept threshold combines very well with CTBV, providing a cheap solution emphasizing power efficiency and high speed through its binary quantization and targeted dynamic range. One essential advantage of swept threshold sampling compared to multi-bit systems is that it required only a 1-bit quantizer, which simplifies the design and provides good inherent linearity of the system [8].

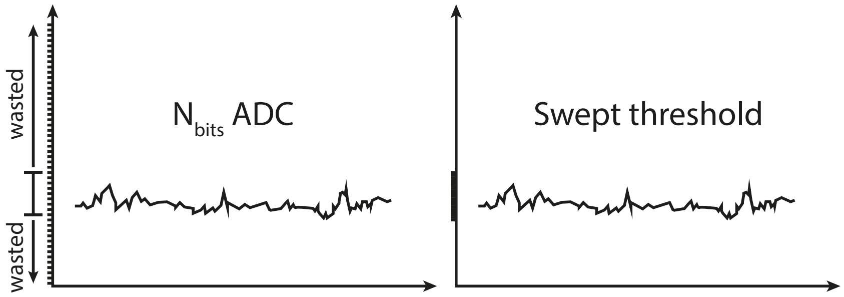


Figure 2.4: Multi-bit A/D conversion vs. swept threshold. Swept threshold allows us to adjust the sweep range and threshold steps depending on the input signal. This is yields an increased SNR while being more power efficient for weak signals. It is also faster and reduces the complexity of the sampling system.

In the backscattered signal, a significant relative level of uncorrelated white noise is expected. For a given threshold level this noise will lead to variations in the thresholded pulse widths and sometimes result in disappearing pulses. This might seem like a bad thing, but in fact, the white noise present can contribute in recovering weak signals which would otherwise go undetected. Signals below a given threshold level would by themselves pass unsensed, but with the added noise they will sometimes exceed the threshold and be picked up by the sampler. This behavior is referred to as *stochastic resonance* [9].

A full signal reconstruction could theoretically be achieved by using multiple quantizers similar to a flash ADC. 2^N quantizers with individual threshold levels at linear intervals would provide a N_{bits} representation of the signal in a single sampling. However, to generate a 16-bits representation of the signal a total of 65535

[2] Threshold level is set by an external high-resolution DAC

quantizers, each with individual threshold levels is required on-chip, which is not feasible [10]. By using a single quantizer, and repeatedly sampling the signal a total of $2^{N_{bits}}$ times, each time with a unique threshold level, and then accumulating these results with counters, we achieve the same without the need of multiple quantizers. Besides, this method of sampling also has the benefit that a few bit errors now and then will not cause any significant errors. By integrating over multiple sweeps, the system becomes more robust, whereas, in a more conventional sampling system, e.g., a flash ADC, any such bit errors *could* have severe consequences on the results. The level of noise in the signal will determine how fine the threshold steps (ΔV_{th}) should be in order to recover the signal with adequate noise reduction. With noise levels greater than ΔV_{th}, one can make the steps finer (reduce ΔV_{th}). However, this is not always an option, e.g., limitations in DAC that sets threshold levels. Instead, stochastic resonance can be utilized with swept threshold by performing multiple samplings at each threshold level. This is particularly useful in recovering weak signals where the noise is in the same order of magnitude as the signal, which is the expected scenario for the pulsed LIDAR.

2.5 System architecture

As presented so far in this chapter, the basis for realizing a high-resolution noninvasive biomarker sensor has been laid. A quick summary of the previous sections:

- Transcutaneous sensing with targeted NIR light provides high depth of penetration in body tissue, but great absorption in tissue severely weakens the backscattered signal. With complex light-tissue interactions, a solution in time-of-flight, pulsed LIDAR, holds potential in addressing the challenges regarding selectivity. By measuring scattering at specific depths, selectivity may be significantly improved.

- High resolution is achieved with high sampling rates. 1-bit signal processing takes advantage of the inherent benefits of CMOS; speed and simplicity. Substituting clock synchronization with delay-based timing enables a system to operate in continuous-time with an operating speed limited by the inherent gate delay of the technology (1mm requires 100GS/s = 10ps).

- The coarse quantization and high potential rates of CTBV may help in increasing sensitivity as well. Signal averaging can, through repetition, manage to recover weak signals buried in (uncorrelated) noise. With very high sampling rates we can safely repeat measurements at frequencies in the MHz range (high PRF). The process in the body are very slow in comparison, so there will be practically no change over the course of a sampling period (blood glucose level show no significant change within a fraction of a second).

- The swept threshold technique allows us to adjust the dynamic range. The range and number of threshold levels (the step between) them decide the dynamic range of the sampler.

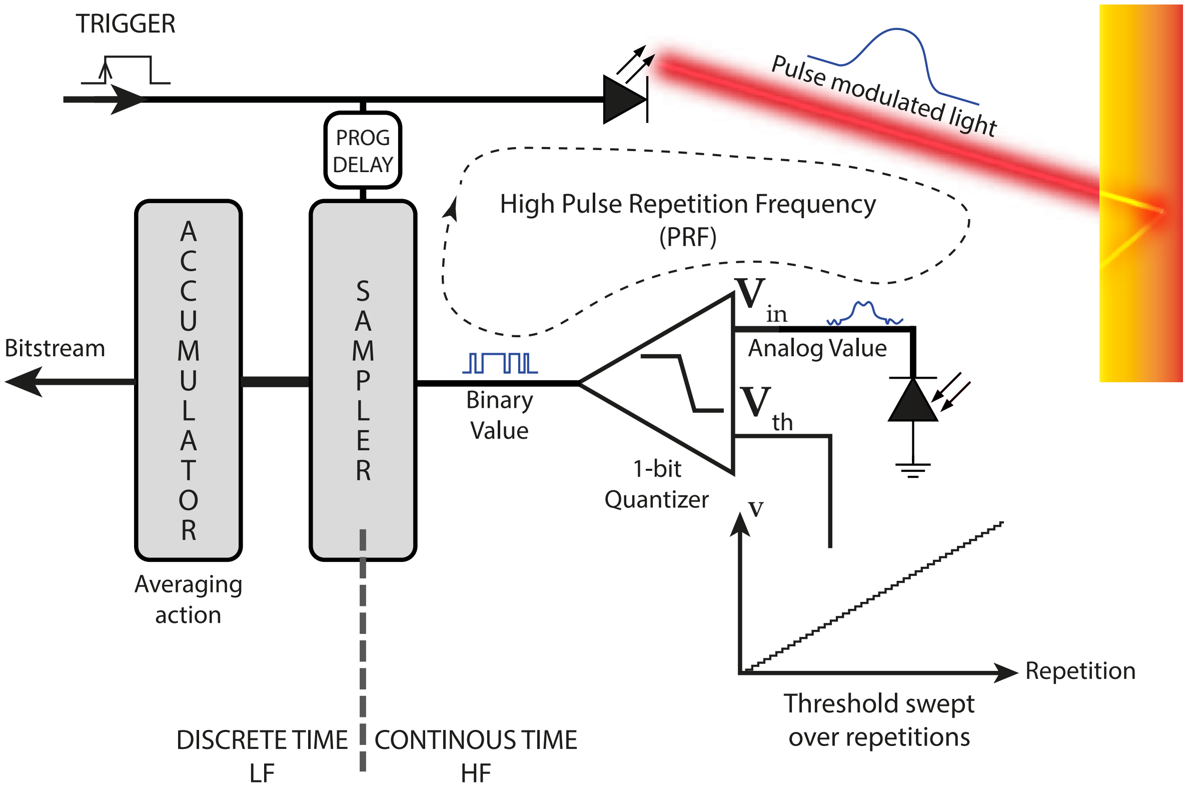


Figure 2.5: System architecture

Figure 2.5 illustrates the system architecture of the proposed pulsed LIDAR system. Conceptually we may consider the proposed LIDAR to be a modified radar with a RF-to-light transducer (laser-diode or LED) on the output and one or more light-to-RF transducer(s) (photodiode(s)) on the input. The weak echo pulses from modulated light directed into the body are sensed by photosensor(s) on the die. The signal is then thresholded and quantized, leaving it digital, yet still continuous. Digital samplers, triggered in quick succession, now performs the high-speed sampling of the signal. After sampling, accumulators connected to each sampler are incremented if a "1" was sampled. This procedure is repeated while sweeping the threshold level across the range of the input signal, resulting in a full readout of the received signal. This method of "sample and accumulate" through the range of threshold levels appear as an equivalent A/D conversion.

The down-range resolution of backscattered signals from static targets is proportional to the signal bandwidth. Since the signal bandwidth is inversely proportional to the signal duration in the time domain, a short light pulse will provide the highest depth resolution. Fortunately, wide signal bandwiths are allowed for modulated light (modulated light does not have to comply with the strict RF regulations of FCC[3] and ETSI[4]). However, the modulation bandwidth will become limited by the bandwidth limitation of the transducers. In this thesis, our primary objective is designing the high-speed sampler. It should be made clear that neither the quantizer or the photosensor(s) are within the scope of this thesis, and will therefore not be subject of any further discussion beyond this paragraph. I will limit myself to illuminating the necessary characteristics of the photosensor and the main features of the quantizer. With weak high-frequency backscatter, the photosensor(s) must be fast while also providing high responsivity for the wavelength range of interest. Designing photodiodes in standard CMOS (silicon substrate) that have both these characteristics at the same time come with its own set of technical difficulties. Suffice it to say that when confined to standard CMOS one of the only ways of increasing the speed of a photodiode is to reduce its junction capacitance, i.e., the active area. This increased speed comes at the expense of lower sensitivity, which means a loss in responsivity. Also, a smaller photodiode will have a higher noise floor, which can become a limiting factor considering the low light levels. The quantizer, developed by Novelda [11], converts the analog RF input signal to the CTBV domain in continuous time, able to react to minute changes in the input signal at high frequencies, with fine-grained control over threshold levels. It has shown good performance in similar systems, such as the successfully implemented swept threshold radar by

[3] Federal Communications Commission
[4] European Telecommunications Standards Institute

Hjortland and Lande [6].

In short, the sampling system can be subdivided into four individual subsystems:

- *Timing - Delay line(s)*: Each unit delay of the dealy line consists of double inverters.
- *1-bit Samplers - D flip-flops:* Rising-edge triggered master-slave D flip-flops.
- *Accumulators - Counters:* Asynchronous counters that increment with pulses applied to the counter's "CLK" input (sampler output).
- *Readout - Shift registers*: Parallel in/Serial out (PISO) shift registers makes counter data accessible over SPI.

The implemented D flip-flops, counters and shift registers are detailed in chapter 3. Both the counters and shift registers are trivial in their design and since they are considered a part of the low-frequency digital domain, and require minimum attention at this point. The D flip-flop, however, can be viewed as the link between the high-frequency and low-frequency domain, and as such might prove to have a definitive impact the sampler performance due to a phenomenon referred to as *metastability*. This will be discussed in section 3.3. A cascade of delay elements makes up a delay line, which provides a history of an input signal. An applied trigger pulse will appear to travel within the line appearing at discrete points in time determined by the individual delay of each element. By using a delay line to trigger a sequence of samplers, tapped at these intermediate states, all with the thresholder output as their input, we have a sampling system which performs down-conversion from tens of gigahertz sampling rate to rates much more manageable by the accompanying digital logic (counters). With the explicit correlation between sampling rate and the equivalent spatial resolution, the delay lines will be of the utmost importance as they will determine the sampling rate of the sampler, and should be designed with elements at a minimum delay. However, with a small middle ground between success and failure, they will introduce significant design challenges.

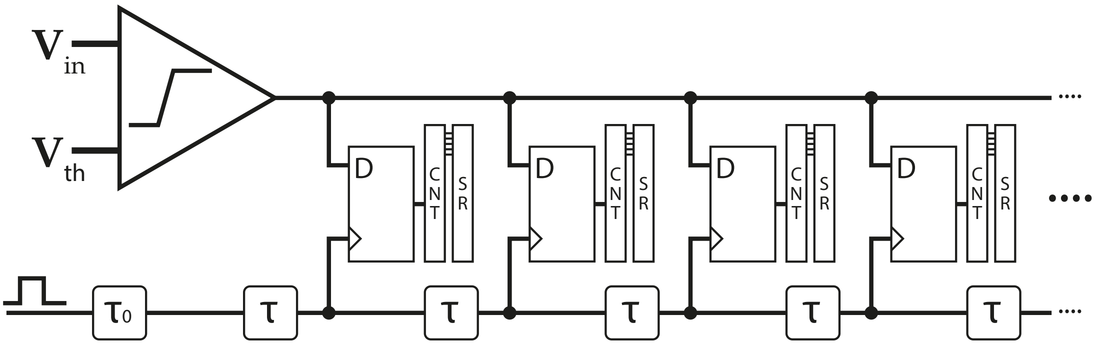


Figure 2.6: Conceptual block diagram of swept threshold sampler

$$f_T \approx \frac{g_m}{2\pi C_L} \tag{2.8}$$

$$g_m = \frac{2I_D}{V_{ov}}\ ,\ \ V_{ov} = V_{GS} - V_{th} \tag{2.9}$$

$$I_D = \frac{\mu C_{ox}}{2}\frac{W}{L}V_{ov}^{\alpha}\ ,\ \ \alpha \in [1,2] \tag{2.10}$$

In the following paragraphs, we will investigate how short gate delays we can achieve in standard 90nm CMOS technology, how we can use this to reach our targeted sampling rate and touch on the challenges introduced in this design. Equation 2.8 refers to the transition frequency of a transistor f_T. It is defined as the frequency at which the current gain ($\frac{I_{DS}}{I_{GS}}$) becomes unity, i.e., it represents the frequency where it transitions from an amplifier to an attenuator [12, p. 291]. It is determined by the relationship between transconductance (g_m) of the transistor, and the total capacitive load (C_L). Equation 2.9 shows that g_m is proportional to the drain current, I_D. Intuitively this makes sense, since the higher the current, the faster it will charge its load capacitance. Equation 2.10 describes how I_D relates to the parameters of the transistor. μ is the carrier mobility, C_{ox}

the gate oxide capacitance, W and L are the dimensions of the gate, while V_{ov} is the gate overdrive voltage. For a given process node, C_{ox} is assumed fixed, while the mobility μ, a fickle parameter, highly dependent on doping concentration, can also be considered a process dependent parameter. This means that for a given supply voltage, the designer is left only the device dimensions (primarily W) to adjust the gate delay of a transistor. Note that the majority of C_L consists of the gate capacitance ($C_{ox}\frac{W}{L}$) of the following gate(s), with additional contributions from $C_{wire}+C_{diff}+C_{parasitic}$. This means that when scaling one gate to increase drive, the previous stage will see an increased load, increasing its delay. Increasing the gate overdrive voltage, i.e., increasing the difference between V_{GS} and V_{th}, can help to reduce the delay. This can be achieved by either increasing the supply voltage (within the constraints of the technology), which comes at the cost of greater power consumption ($P_{dyn} \propto C_L VDD^2$), or as we will explore in subsection 3.2.2, through back-gate tuning which allows adjustment of the threshold voltage and grants some degree of delay adjustment.

Previous publications [13], [14] indicate a nominal transition frequency for a single inverter approaching $100GHz$, or $t_{pd} = 10ps$. However, considering parasitics and the added load on the output, a more realistic assumption is that a single inverter can operate with a propagation delay of around $15ps$. In other words, it is feasible that a delay element consisting of two inverters can be designed to approach a delay as low as $30ps$ reliably. Based on this lower limit in delay, a single delay line will allow for an effective sampling rate of approximately $33GHz$ ($\frac{1}{30ps}$), as demonstrated by Hjortland and Lande [6]. However, the targeted depth resolution of 1mm equivalates to a sampling rate of $100GS/s$, which means that the 1-bit samplers must be triggered at $10ps$ intervals. It becomes apparent that if we are to improve speeds any further, we will require parallelization to overcome this fundamental limitation in the technology. As mentioned in section 2.2, to further increase the time resolution in CTBV systems the relative difference between two delay elements can be exploited. By delaying signal A with for example $50ps$, and signal B with $60ps$ we obtain a relative delay of only $10ps$. This relative difference between delays can be used to create a timing configuration consisting of three parallel delay lines, in which each line is initiated with a relative time offset of $10ps$ to its neighbor(s). This approach of interleaving delay lines is illustrated in Figure 2.7 where, in this case, three delay lines are interleaved. This technique effectively triples the sampling rate and depth resolution, and is based on the premise that each initial delay differ from their neighbor with exactly $10ps$ and that the subsequent elements in each line all accurately generate a delay of $30ps$ in order to maintain a linear behavior and yield an effective sampling rate of $100GS/s$. This will, however, not be the case. Since the conception of CMOS, stochastic variations in process parameters has induced a random mismatch in threshold voltage and current factor (β) between identical devices. These unavoidable variations become more apparent as transistor sizes decrease, and will become a determining factor in the characterization of the sampling system. Device mismatch will limit the timing precision of the trigger pulse propagating through the delay lines and introduces restrictions on both speed and on how deep the delay lines can be designed.

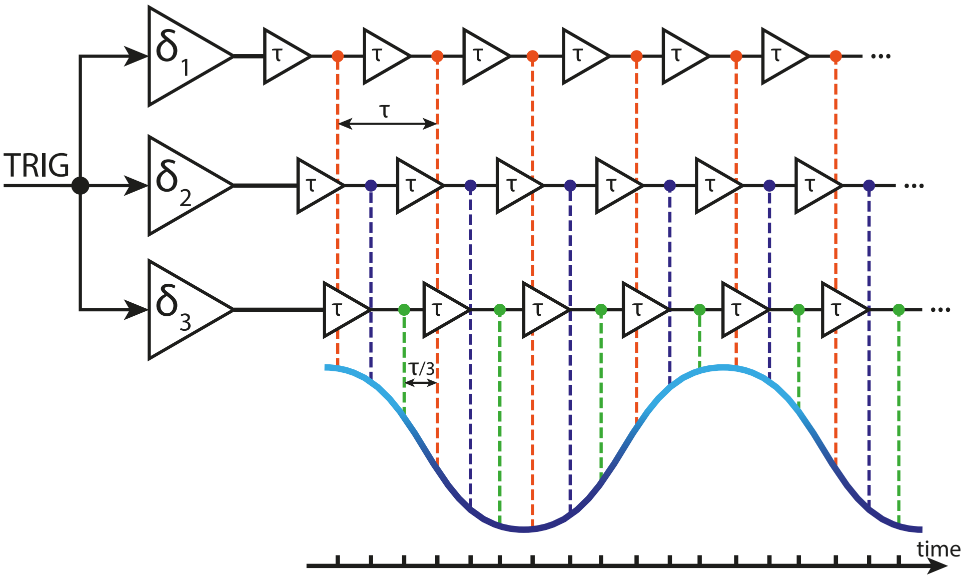


Figure 2.7: Delay line interleaving. $\delta_2 = \delta_1 + 10ps$, $\delta_3 = \delta_2 + 10ps$, $\tau = 30ps$

2.6 Static Variations

2.6.1 Process variation

As highlighted in the previous section; sufficiently fast delay elements are obtained by designing transistors with short and wide channels which reduce their capacitive load and maximize their drive potential, yielding fast switching transistors. When we are interleaving lines it is vital that all delay elements produce accurate delays. However, all transistors, particularly small ones, will suffer from the influence of process variations. These inescapable variations, originating from random atomic-level fluctuations and limitations in manufacturing, cause behavioral differences *between* devices, which means that no two CMOS devices are ever precisely alike. At smaller dimensions and lower supply voltages process variations are becoming increasingly relevant, since their contribution becomes a larger relative percentage of the device. Despite strides in the technologies, process variation remains a significant challenge in achieving high-speed, low voltage CMOS systems.

Process variation manifests across two kinds: interdie (die-to-die) and intradie (within-die) variations [15], [16] (in reality, also lot-to-lot and wafer-to-wafer, which mainly concerns high volume production, so we discard the considerations regarding such variations). Interdie variations, caused by systematic effects like process gradients on the wafer, are independent of device sizes. Interdie variations act globally on the entire die so that each device on the die experiences the same random shift in the mean value of device parameters. Intradie variations, however, affect each device within the same die in a different way. Intradie variations, commonly called *device mismatch*, can be categorized as random or systematic. The random component has a Gaussian distribution and shows no correlation across devices, which means that it cannot be predicted beyond statistical estimations. The systematic fluctuations exhibit spatial correlation and thus, lead to minimal variation for devices in close proximity. The relative effect on the mismatch due to device distance is only significant for large area devices with considerable spacing [16] and is therefore not considered in this thesis. Our definition of mismatch excludes all systematic variations and offsets of the absolute value of parameters caused during manufacturing by electrical, lithographic or timing differences.

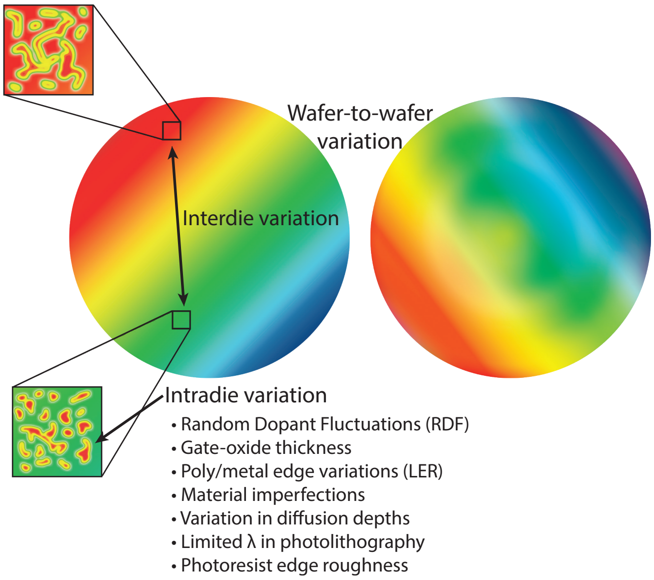

Wafer-to-wafer
variation
Interdie variation
Intradie variation
• Random Dopant Fluctuations (RDF)
• Gate-oxide thickness
• Poly/metal edge variations (LER)
• Material imperfections
• Variation in diffusion depths
• Limited λ in photolithography
• Photoresist edge roughness

Figure 2.8: Process variations

In 1989 Marcel Pelgrom established that variations in threshold voltage and current factor (β) are inversely proportional to the square root of the transistor area (Equation 2.11 and 2.12) [16]. This relationship, which has been shown to hold up quite well in modern CMOS technologies as well [17], intuitively means that the influence from mismatch on V_{th} and β will decrease with increased device sizes since the parameter variations will "average" over a greater distance or area. Note that V_{th} and β are not considered process parameters, but rather a result of several parameters, such as gate oxide thickness and substrate doping concentration. This shows that the local variation of such process parameters, p, are themselves dependant on the transistor dimensions (2.13). Furthermore, his experiments showed that threshold voltage variations are the dominant sources of drain-source current mismatch, which directly affects the gate-delay.

$$\sigma_{V_{th}} = \frac{A_{V_{th}}}{\sqrt{WL}} \tag{2.11}$$

$$\sigma_\beta = \frac{A_\beta}{\sqrt{WL}} \tag{2.12}$$

$$\sigma_p \propto \frac{1}{\sqrt{LW}} \tag{2.13}$$

where W is the gate-width and L is the gate-length, and the proportionality constants $A_{V_{th}}$ and A_β are technology-dependent.

$$V_{th_N} = V_{th_0} + \gamma\left(\sqrt{|V_{SB} + 2\phi_F|} - \sqrt{|2\phi_F|}\right) \tag{2.14}$$

$$\text{where } \gamma = (t_{ox}/\epsilon_{ox})\sqrt{2q\epsilon_{Si}N_A} \tag{2.15}$$

$$\text{and } \phi_F \text{ is the surface potential} \tag{2.16}$$

The body effect equation (2.14) which describes the resulting threshold voltage when a substrate bias is applied to (in this case) a NMOS transistor (body- or back-gate-effect), helps us illuminate how parameter variations have a direct effect on the electrical properties of a transistor. As the equation shows, V_{th} is proportional to the body effect parameter (γ) which includes the oxide thickness (t_{ox}) and doping concentration (N_A). Intuitively any variations in these parameters would result in variations in V_{th}. Mizuno et al. [18] established that depletion layer charge fluctuation due to random dopant fluctuation (RDF) is the primary cause of V_{th} variation. In modern technologies RDF has been shown to have a more substantial impact since the number of dopants is substantially fewer [19]. They also showed that V_{th} fluctuations as a result of variations in the oxide thickness, although small, decreased linearly with decreasing oxide thickness (Equation 2.17). For a long time, as technologies scaled down, mismatch consistently improved due to a steady decrease in gate oxide thickness and better channel doping control which resulted in lower threshold voltages. In the last decade, however, the oxide thickness and threshold voltage scaling have significantly slowed down, as some parameters are approaching fundamental limits, both physical and to avoid unwanted effects such as time-dependent gate oxide breakdown[5].

Due to the random nature of ion implantation, dopant diffusion and other processes involved in the doping of silicon, the number of channel dopants exhibits a statistical variation given by the Gaussian function [20]. With a discrete and random number of dopant atoms in the channel, there emerges an inherent spreading of various transistor parameters. Equation 2.17 shows how $\sigma_{V_{th}}$ depends on t_{ox}, N_A and the device area. It has been shown by later publications, that the analysis done by Mizuno is somewhat limited as it does not take into account all process contributions, such as S/D dopant fluctuation and substrate bias [20], but within the scope of this thesis, it provides sufficient insight into the parameters influencing $\sigma_{V_{th}}$.

$$\sigma_{V_{th}} = \left(\frac{\sqrt[4]{4q^3\epsilon_{Si}\phi_B}}{2}\right) \times \frac{t_{ox}}{\epsilon_{ox}} \times \frac{\sqrt[4]{N_A}}{\sqrt{W_{eff}L_{eff}}} \tag{2.17}$$

where $\phi_B = 2k_BT\ln(N_A/n_i)$ (k_B: Boltzmann's constant, T: temperature, n_i: carrier concentration, q: elementary charge), and ϵ_{Si} and ϵ_{ox} are the permittivity of silicon and oxide. L_{eff} & W_{eff} are used to not include the offsets in the device area introduced by L/W variation.

[5] A conducting path is formed through the gate oxide to substrate due to electron tunneling.

Process Parameters	Electrical Parameters
Flatband voltage (V_{fb})	Drain-source current (I_{DS})
Mobility (μ)	Input voltage (V_{GS})
Dopant concentration (N_A)	Trans-conductance (g_m)
Length offset (ΔL)	Output Conductance (g_0)
Width offset (ΔW)	
Short Channel Effect (V_{tl})	
Narrow Width Effect (V_{tw})	
Gate Oxide Thickness (t_{ox})	
S/D Sheet Resistance (ρ_{sh})	

Table 2.2: [21] Process parameters manifest changes in the electrical behavior of a device, whereas the electrical parameters are those parameters usually of interest to the designer. V_{th} is highly dependent on N_A, which is not a finite number, as it depends on L, W, and the body bias. For this reason V_{th} does not belong in either category.

The process parameters that influence mismatch are fundamentally bound to the technology itself and are thus not variables that the designer can tune. During design, the circuit designer is left with only the device dimensions and layout to control the matching. However, during operation, the gate overdrive voltage, defined as the difference between the gate-source voltage (V_{GS}) and the threshold voltage (V_{th}), has a substantial influence on the contribution from threshold voltage variations. The resulting drain current modulation for a given V_{th} shift will be diminished by increasing V_{ov}. This becomes evident when studying the generic Sakurai model equation for drain-source current in saturation (2.18) [22]. For digital logic, V_{GS} in the equation can be considered to be VDD.

$$I_{DS,sat} = \frac{\mu C_{ox}}{2}\frac{W}{L}(\overbrace{\underbrace{V_{GS}}_{\text{VDD}} - V_{th}}^{V_{ov}})^{\alpha} \tag{2.18}$$

where α is the velocity saturation index ranging between 1 and 2; 2 for long channel and close to 1 for short channel (sub-100nm) technologies. Considering the linear relationship between saturation current and V_{ov}, short channel devices are less sensitive to V_{th} (and VDD) variation than long channel devices, but it still remains significant. The supply voltage is limited by technology specifications and is not a variable that can be adjusted extensively to increase V_{ov}. Applying excessive voltages can cause a breakdown of the internal junctions of a transistor, which leads to an exponential increase in current, and high power dissipation. The trend of reducing supply voltages has made the impact of V_{th} mismatch more severe, however, many technologies provide low threshold transistors (LVT) that will operate with greater V_{ov} (at the cost of increased leakage current). At a fixed power supply voltage, low threshold devices become less sensitive to variations in threshold voltage (or VDD).

2.6.2 Mismatch in the delay lines

In most systems, there will be paths that are more sensitive to mismatch than others. These *critical paths* tend to become bottlenecks and often require particular attention. In the presence of mismatch, all devices will exhibit a random tendency to stray from their expected characteristics, such as the propagation delay. Any two (seemingly) identical paths will produce different, uncorrelated delays. The impact of mismatch on the propagation delay of a digital logic path can be investigated by studying a string of n identical logic gates, each with a mean gate delay of μ_g and a standard deviation of σ_g. The total mean path delay, μ_{path}, will display a linear relationship with n, while the total standard deviation of the path delay, σ_{path}, increases only proportional to $\sqrt{n}$ [19]. This means that the *relative* delay variation of the path, its coefficient of variation (CV), decreases proportionally to the inverse square root of n. Note that it is the relative uncertainty that decreases with the path length, not the absolute standard deviation. Even though the delay of a long path will have a higher standard deviation than a short path, its standard deviation relative to the mean will be lower, i.e., its delay variability is lower (Figure 2.9).

$$\mu_{path} = n \cdot \mu_g \tag{2.19}$$

$$\sigma_{path} = \sqrt{n} \cdot \sigma_g \tag{2.20}$$

$$CV = \frac{\sigma_{path}}{\mu_{path}} = \frac{\sqrt{n}}{n}\frac{\sigma_g}{\mu_g} = \frac{1}{\sqrt{n}}\frac{\sigma_g}{\mu_g} \tag{2.21}$$

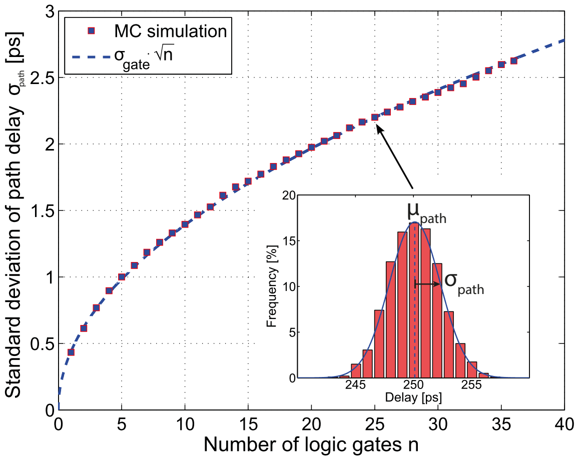


Figure 2.9: Effect of mismatch on critical path delay [19]

This means that long paths exhibit an "averaging" that tends to generate a total delay that is closer to the expected mean value, while shorter paths, generate a shorter delay, but display a greater tendency to differ from its expected delay. In our delay lines, it is not the total propagation delay of the line(s) that is of interest, but rather the specific delay of each delay element that makes up the line. Remember, the samplers tap the trigger pulse as it appears at each intermediate stage of the line(s). Accordingly, this means that we must consider the delay element our *critical path*. With a critical path of only two minimum length inverters, we expect a high coefficient of variation. The high relative delay mismatch in each delay element is expected to give rise to a more nonuniform propagation of the trigger pulse, which will result in inequidistant sampling. The variation at one stage in the delay line will affect all subsequent stages. In other words, the deviation in the nth stage will include the accumulated contribution of delay mismatch from all the preceding stages. Intuitively, this accumulation of mismatch leads to an increasing deterioration as the number of cascaded delay elements increase. As previously mentioned, delay line interleaving will be necessary to increase the effective sampling rate beyond the limitations set by the shortest gate delay of the technology. An important observation to be made at this point: When considering the accumulative deterioration of certainty in delay present in each delay line in combination with interleaving we see that there exists a possibility that samplers can trigger in non-sequential order. The probability of this happening increases the further into the line(s) the trigger pulse travels, meaning that there will be a limitation on the depth of the lines, determined by the degree of mismatch in the delay elements. The problem is illustrated in Figure 2.10.

Remark: This deterioration will also be impacted by jitter (section 2.7).

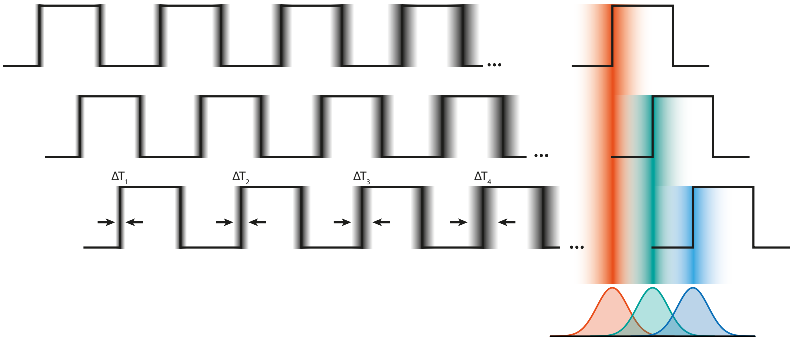


Figure 2.10: Accumulative deterioration of timing accuracy due to mismatch in the delay lines.

In section 3.2 we will perform Monte Carlo simulations (MC) on the delay elements to help "predict" the influence of mismatch on the delay line characteristics. MC is a statistical analysis that assists us in determining the influence of mismatch by performing simulations over a wide range of randomly chosen device parameters (corners), in accordance to the mismatch model of the technology. The results provide a statistical prediction on how the collective impact of process variations accumulate and affect the delay line performance, which will assist us in the design effort.

2.6.3 Summary

- Mismatch is caused during manufacturing by statistical variations in the process parameters that determine the behavior of devices. The random and uncorrelated nature of mismatch means that no two devices are ever identical.
- The parameters are fundamentally bound to the technology. For a given technology, the impact of mismatch can only be mitigated only with increased device area and through increasing the gate overdrive voltage. During operation, compensation for mismatch can be achieved by back gate tuning devices.
- Path delay variations become more prominent for smaller dimensions and reduced supply voltages. The relative delay variations in a path also display a dependence on its depth.
- The delay elements that make up the delay lines have been identified as *critical paths*, expected to have a high relative delay variation due to their minimum size and low depth. The accumulation of mismatch in the delay line places a constraint on the depth of the delay lines.

2.7 Environmental Variations

While process variations originate from the limitations of manufacturing, environmental variations are the result of the surrounding environment of the chip during its operation. These variations incorporate temperature variations, power supply variations and variations in switching activity which result in a varying performance. The small leakage currents of CMOS devices produce a minimal static power consumption (P_{stat}). However, when switching at high frequencies, the dynamic power consumption (P_{dyn}) will contribute significantly to the overall power consumption of a system. The dynamic power consumption is the sum of the transient (P_T) and the capacitive-load (P_L) power consumption. As a device switches from one state to another, there is a transient current spike that originates from the charging of the internal nodes and a brief current that flows from VDD to GND while both N- and PMOS are on at the same time. Considering the fast transitioning delay elements, this *"through current"* is considered negligible compared to the charging current. Further power is consumed in the charging/discharging the applied load capacitance(s) of the device [23].

$$P_{tot} = P_{stat} + P_{dyn} \tag{2.22}$$

$$P_{stat} = I_{leak} \cdot VDD \tag{2.23}$$

$$P_{dyn} = P_T + P_L \tag{2.24}$$

$$P_T = C_{PD} \cdot VDD^2 \cdot f \tag{2.25}$$

$$P_L = \sum (C_{Ln} \cdot f) \cdot VDD^2 \tag{2.26}$$

where

I_{leak} is the leakage current

C_{PD} is the internal capacitance of the device

C_{Ln} is the capacitance of load n (fanout)

With a variable switching activity of the chip, the current demands on the supply voltage will consequently vary. With parasitics in the power delivery system, this changing power requirement results in a fluctuating supply voltage, i.e., power-supply noise. Analog devices are particularly vulnerable to power-supply noise since they often rely on stable bias points. In digital logic, the situation is not as severe since the noise margins for false transitions are significantly wider. However, the supply noise will affect their timing properties. In our delay lines, this noise will introduce a timing jitter which translates to an increased variation in the sampling instants [24]. The power-supply noise causes variations in the drive strength of the transistors and modulates their delay (see Equation 2.8). The uncorrelated variation in behavior between transistors caused by mismatch means that each device will be affected differently by this noise. In section 3.8 we will discuss how the power-supply noise is reduced in our design.

Furthermore, the generated heat during operation leads to fluctuating gate delays. This can be attributed to both a temperature induced threshold voltage variation and carrier mobility fluctuation [25]. Equation 2.27 shows how the surface potential (ϕ_F) is proportional to the absolute temperature T. Remembering the body effect equation (Equation 2.14) this means that the threshold voltage reduces as the temperature increases, which in turn improves the gate delay. Note that as the generated heat lowers the threshold voltage of transistors, leakage currents are increased. These increased leakage currents lead to higher static power consumption, which in turn contributes to increased temperature. Observe how this positive feedback can cause failure if strong enough.

$$\phi_F = \left(\frac{kT}{q}\right) \ln\left(\frac{N_A}{n_i}\right) \tag{2.27}$$

where k is Boltzmann's constant, q is the elementary charge, and N_A and n_i are doping parameters.

Whereas the threshold voltage decreases with temperature, the carrier mobility displays an inversely proportional response to temperature, as shown by Equation 2.28. This means that the contributed gate delay modulation from mobility is therefore negative for an increased temperature.

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-k_\mu} \tag{2.28}$$

where T_r is the room temperature and k_μ is a fitting parameter ~ 1.5

It can be summarized that environmental variations can have a notable effect on the precision and delay resolution of delay lines during operation. These variations are not random since they depend on the workload of the system and is therefore time-dependent. The scenario which results in the worst-case voltage supply drop can occur at any time and is accordingly extremely difficult to predict. Likewise is the temperature dependent on a multitude of factors such as the ambient temperature and how well the chip dissipates heat through conduction, convection and radiation.

Chapter 3

Implementation

In this chapter, we will describe how the proposed architecture has been implemented in CMOS. As we introduce the various sub-systems, relevant analysis and discussion are presented to reveal the reasoning behind the design.

3.1 Sampling System

Figure 3.1 shows a block diagram of the sampling system. Additional supporting circuitry has been deliberately excluded at this stage to provide a clear system overview. However, as additional circuits become relevant, they will be introduced and discussed in detail throughout this chapter. The core subsystems are color-coded as such: **TIMING**, **SAMPLING**, **INTEGRATION**, **READOUT**.

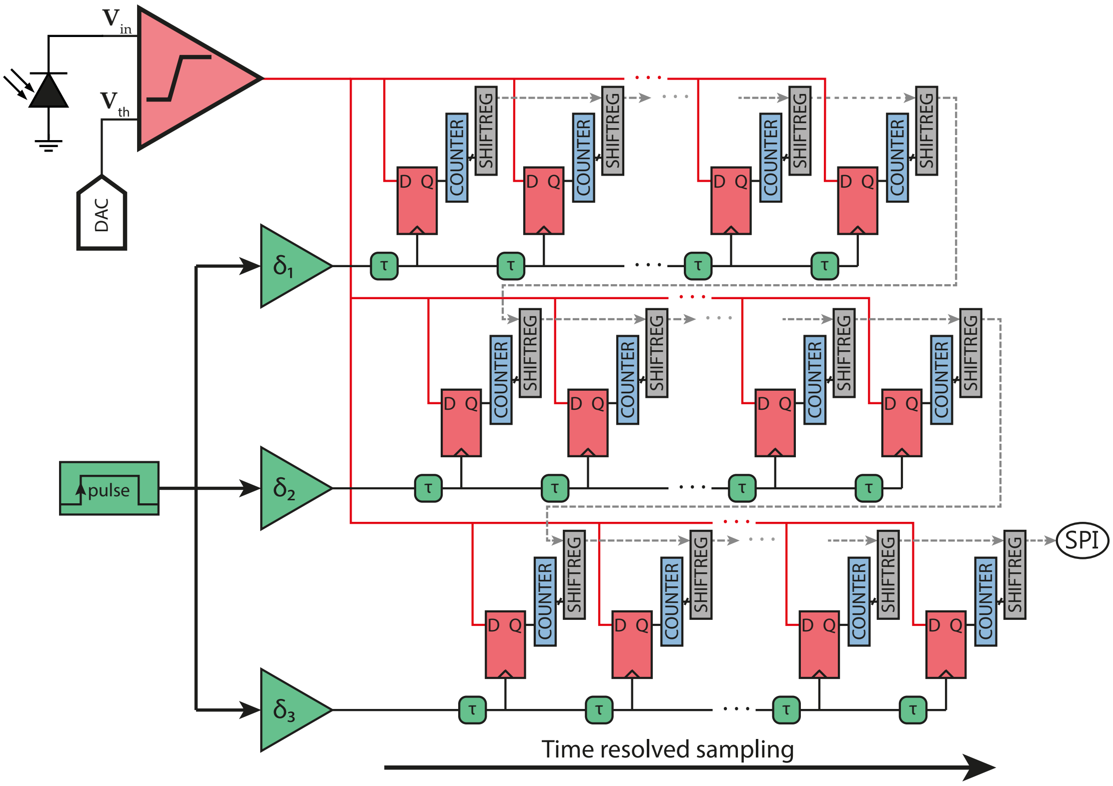


Figure 3.1: Simplified block diagram of the sampling system, $\delta_2 = \delta_1 + 10ps$, $\delta_3 = \delta_2 + 10ps$, $\tau = 30ps$.

3.2 Delay lines

The "spine" of the system can be considered to be the three interleaved delay lines. Each line consists of 16 cascaded delay elements of double inverters designed to have a unit delay of $\tau = 30ps$. Preceding each line is an initial delay element whose purpose is to create a relative offset of $\delta\tau_o = 10ps$ to the adjacent line(s). This offset is what enables the interleaved propagation of the trigger signal through the system which, in an ideal scenario will yield an effective sampling rate of $100GS/s$ $(1/10ps)$. For the sampler to operate as intended, it is preferred that the delays are as equally spaced in time as possible. This way the samplers are triggered at regular intervals, generating an *equidistant* sampling sequence. In other words, the samplers trigger in sequential order, with a constant delay of $10ps$. This can only be achieved if each line is initially offset to its neighbor by $10ps$ *and* that the subsequent delay elements do not deviate from $30ps$. To illustrate this more abstractly:

- The term *precision* is used in this context to describe how similar the delay elements are. We relate the term to the degree of variations in the lines, i.e., the standard deviation of the delay elements. Extensive variation (high σ) between the delay elements due to mismatch and/or other effects, i.e., a low precision, will lead to a rapid deterioration of the linearity of trigger pulse occurrences (see subsection 2.6.2). The level of precision is a restricting factor for maximum line depth, which directly relates to how large a window in time we can sample. With severe mismatch in the delay elements, the sampling can become inequidistant or even nonsequential at an early stage in the interleaved system.
- The term *accuracy* is used to indicate how close to the target delay the elements are. The system is designed in such a way that we are highly dependent on the delays reaching or coming close to this target value for the interleaved sampling to be correct. Since a relative initial delay of δ_{τ_0} is used to offset the lines, all subsequent elements must produce a delay of $\tau = \delta_{\tau_0} \cdot 3$ (3 lines) to maintain a uniform temporal spacing between triggered samplers. To achieve the targeted sampling rate of $100GS/s$, δ_{τ_0} must be $10ps$, which means that $\tau = 30ps$. Note that, assuming 100% precision ($\sigma = 0s$), a lack of accuracy, e.g., $\tau = 33ps$, will not cause any deterioration, just a systematic error in the form of a constant inequidistant or non-uniform pulse propagation.

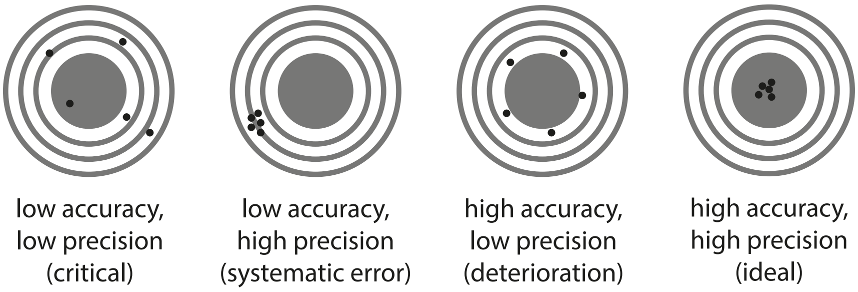


Figure 3.2: Illustration of accuracy vs. precision

In reference to our previous discussions on mismatch (subsection 2.6.2), it becomes clear that when striving to maximize the accuracy of the delay elements, i.e., short and wide transistor channels, there is an inadvertent negative impact on the precision of the lines as a result of an increased presence of mismatch. In this situation of mutually conflicting conditions, the challenge lies in the balancing act between accuracy and precision.

3.2.1 Delay Element

To create these high-resolution delay elements, the transistor lengths are kept at the minimum size $(100nm)$ while the widths are increased to improve their drive strength. The total capacitive load seen at an output is $C_L = C_{self} + C_{wire} + C_{fanout}$, where C_{self} is the combined output junction capacitance of the driving device, C_{wire} is the total interconnect capacitance, and C_{fanout} is the sum of the subsequent gate capacitances. Due to the repeating pattern in the sampler, the load capacitance seen by the elements is close to constant. Thus, apart from the initial custom delays, the elements are identical - designed to have a drive strength which yields a delay of $30ps$. Since it is primarily the rising edge which is of interest to us, the inverters in the delay elements can be designed asymmetrically (to facilitate fast rising edge behavior), which will help to reduce their junction capacitance and consequently both their rising edge delay and the load seen by previous output transistor. As mentioned in section 2.2, this design choice will have the side effect of increasing the width of the trigger pulse as it travels through the lines, but considering the low line depths (16) and relatively long pulse period, it is not expected to cause any complications in this design.

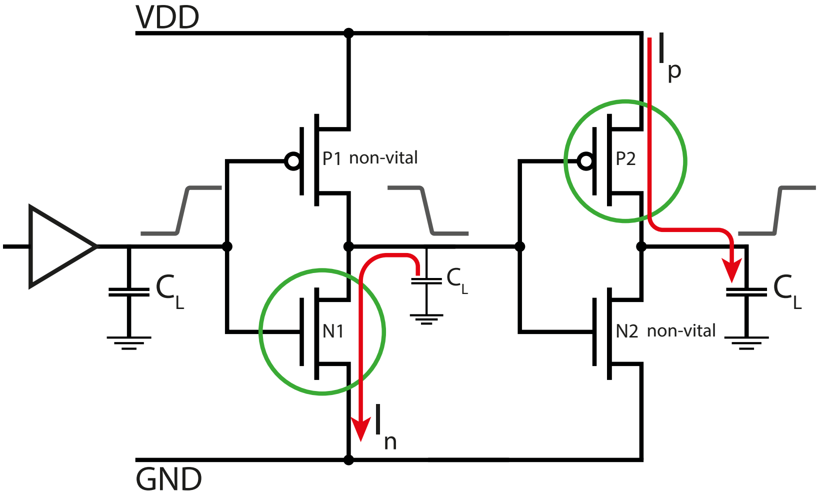


Figure 3.3: By keeping W_{P1} and W_{N2} small the capacitive load is reduced at the expense of slower falling edge transitions.

By using minimum lengths for the transistors, we help to reduce their gate capacitance ($C_G = C_{ox} \cdot W \cdot \mathbf{L} + C_{ov}$). The reduced capacitance, in turn, results in a decreased intrinsic (self/internal) delay (τ_{self}) and also a smaller C_L seen by the previous stage. It should be mentioned that when using short channel transistors, short channel effects become more prominent, which can impact their performance and reliability. First of all, velocity saturation[1] causes short channel devices to not behave according to the square-law model (Equation 2.18). Secondly, as the channel length becomes comparable with the S/D depletion widths, the gate loses some control over the channel. The transistors need to be fairly wide to ensure strong driving capabilities and fast transitions, but the increased widths also lead to higher junction capacitances, which has the negative side-effect of more self-loading (τ_{self} will stay more or less constant since $C_{self} \propto W$). To combat the increase in junction capacitances the transistor fingering or folding technique is applied, where the wide transistor is split into smaller ones connected in parallel. By fingering the transistors, the gate resistance is reduced by a factor N, which makes the transistors turn on and off faster and also has a positive impact on thermal noise in the gate. However, more importantly, the S/D-substrate capacitances are decreased by shared junctions, which further improves the frequency response of the transistors.

[1] velocity saturation: excessive collisions suffered by the carriers cause their velocity to saturate after a critical electric field (V_{gs})

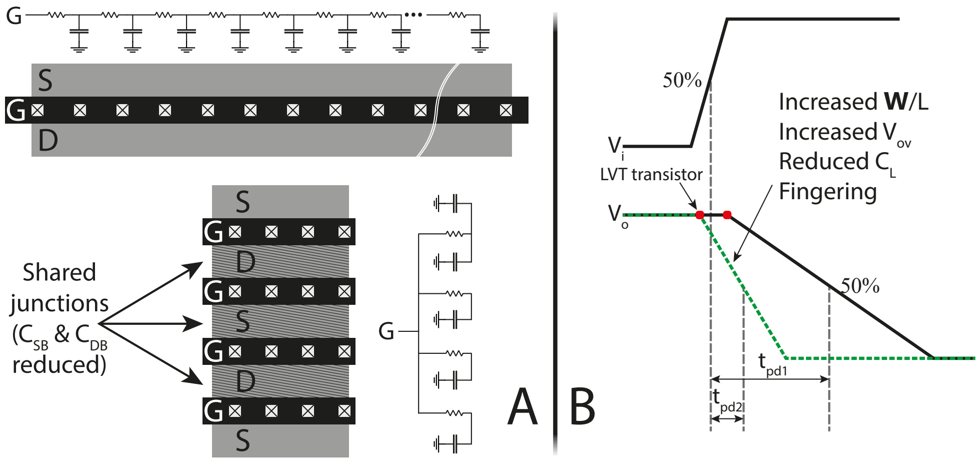


Figure 3.4: **A:** Fingering reduces the gate resistance (R/N) with parallel gates and junction capacitances by shared S/D substrate junctions. In this example, by splitting a wide transistor into four fingers the gate resistance is reduced to $R/4$, while the effective drain and source area is reduced by half and one quarter respectively. **B:** The effects of implementing inverters with high gate overdrive voltage by using low threshold transistors and increased supply voltage, in conjunction with high aspect ratio (W/L), and junction capacitance reduction (fingering).

The delay elements are made up entirely of low threshold voltage transistors (LVT) and are on a separate power supply from the rest of the system. A reduced threshold voltage enables the transistor to trigger at a lower gate voltage and also facilitates a higher overdrive voltage. By having the delay lines on a separate supply, we enable the possibility for further global tuning during operation ($VDD \pm 10\%$) if necessary. Separate supply voltages will also reduce power supply variations on the delay lines, which would otherwise be higher due to the dynamic power consumption in the additional circuitry. As previously discussed, CMOS logic gates become faster with an increased overdrive voltage (and equivalently slower with a decrease). To demonstrate why, a logic gate can be simplified as an RC circuit consisting of the resistance path formed in the channel when a transistor is on, referred to as $R_{DS,on}$, and the load capacitance, C_L. While C_L remains roughly constant, $R_{DS,on}$ display an inversely proportional relationship to $V_{gs} - V_{th}$ (because $R = U/I$), and will consequently decrease with increased overdrive voltage. A more accurate way to think of it is this: as the voltage swing increases, the charging current will also increase to maintain the same speed (remember the capacitor equation, Equation 3.1). However, for a CMOS transistor in saturation, the charging current increases approximately with V_{ov}^{α} (recall the MOSFET equation which states that $I_{DS} \propto V_{ov}^{\alpha}$). For long channel transistors, the velocity saturation index $\alpha = 2$. Short channel transistors have a α much closer to 1 due to velocity saturation, but this effect still contributes to improving switching time ($L = 100nm \rightarrow \alpha \approx 1.3 - 1.4$). A consequence of using LVT cells is that they have a higher static power consumption due to increased leakage currents.

$$\tau_{pd} = RC = \frac{CV}{I} = \frac{C_L VDD}{I_{DS}} \tag{3.1}$$

$$\tau_{pd} \propto \frac{C_L VDD}{\mu C_{ox}(W/L)V_{ov}^{\alpha}} \tag{3.2}$$

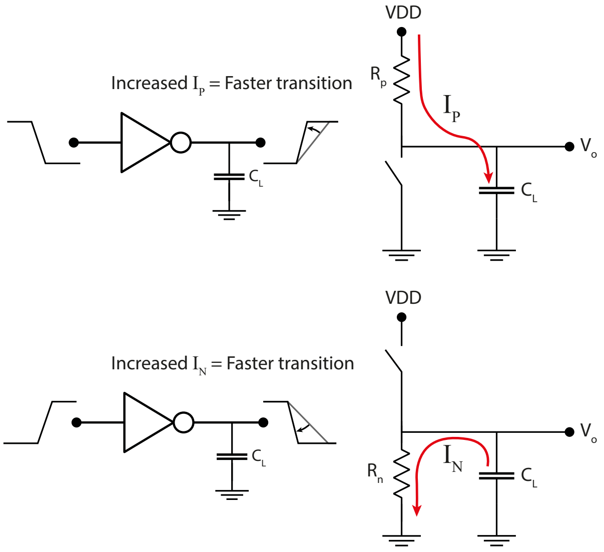

VDD
Increased I_P = Faster transition
R_P
I_P
V_o
C_L
C_L
VDD
Increased I_N = Faster transition
R_n
I_N
V_o
C_L
C_L

Figure 3.5: Illustration: RC circuit

The delay element has been designed with the above points taken into consideration. Sudalaiyandi [26, p. 41] demonstrated that pulse width modulations in a delay line are significantly reduced by designing inverters with a W_P/W_N ratio of 2.1. However, it is not the pulse width that is of importance in this system, but rather the rising edge delay. Therefore we can allow some pulse width modulation if it helps to improve the rising edge delay. The delay elements have been intentionally designed using unbalanced inverters, focusing on minimizing their rising edge delay at the expense of falling edge response. As Table 3.1 shows, the first inverter is significantly better at pull-down than pull-up, whereas the second inverter has a superior pull-up characteristic.

	Inverter 1		**Inverter 2**	
	NMOS	*PMOS*	*NMOS*	*PMOS*
fingers $\boldsymbol{n}$	4	4	4	4
$\boldsymbol{W_n}$	$950nm$	$1.1\mu m$	$525nm$	$1.22\mu m$
$\boldsymbol{W_{tot}}$	$3.8\mu m$	$4.4\mu m$	$2.1\mu m$	$4.88\mu m$
$\boldsymbol{L}$	$100nm$	$100nm$	$100nm$	$100nm$
$\boldsymbol{W/L}$	38	44	21	~49
$\boldsymbol{\frac{Wp}{Wn}}$	1.16		2.32	

Table 3.1: Unit delay element transistor dimensions

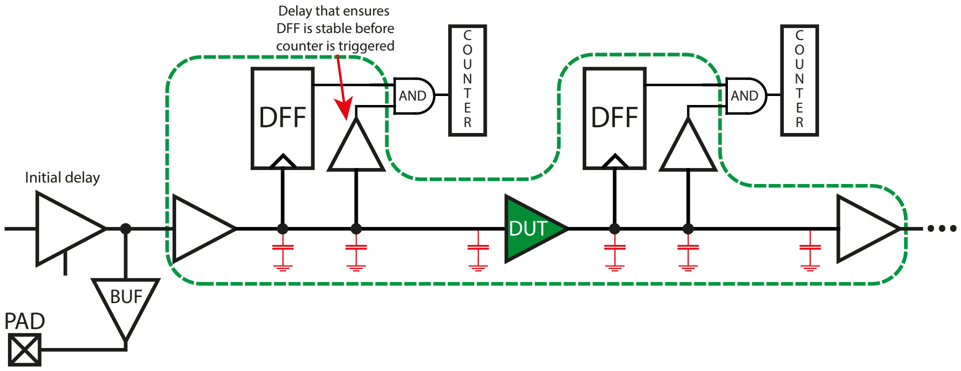


Figure 3.6: Post layout simulations of the delay element involve all gates directly related to the device under test (DUT) to provide an accurate assessment of its performance. This includes the capacitances of both previous and subsequent stages (gate capacitances of the delay element, the D flip-flop and also a customized starved inverter delay element whose purpose is explained in subsection 3.6.1)

Monte Carlo simulations accounting for mismatch displaying the probability distribution of the rising edge delay are shown in Figure 3.7 and Figure 3.8. Observe that with a supply voltage of $VDD = 1.2V$ the delay element is unable to reach $\tau = 30ps$. However, when the supply voltage is increased by 10% to $1.32V$, the element surpasses the target delay. Also, note how the increased supply voltage reduces the relative variation (μ reduced by $\sim 11\%$ while σ reduced by $\sim 22\%$). These observed improvements validate our previous analyses on mismatch and gate delay - that a higher gate overdrive voltage diminishes the influence of mismatch and increases the driving capabilities of the device.

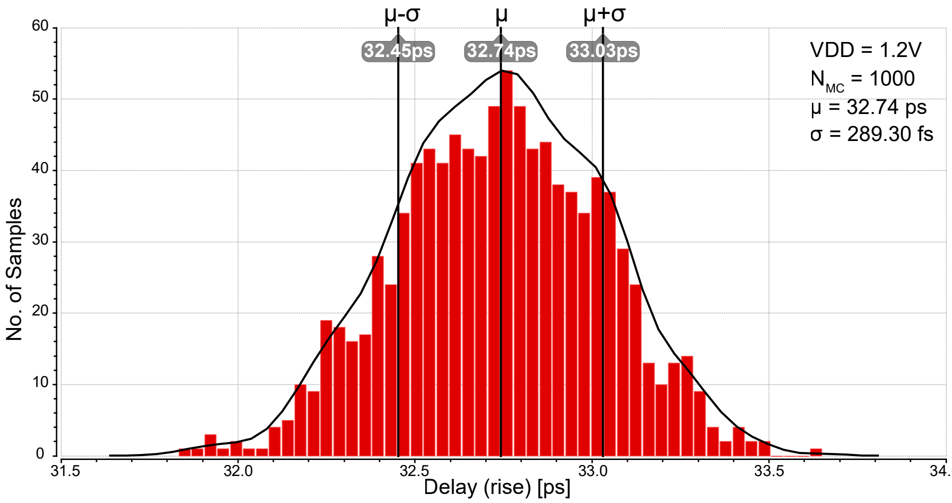


Figure 3.7: Post-layout Monte Carlo simulation of delay element. VDD=1.2

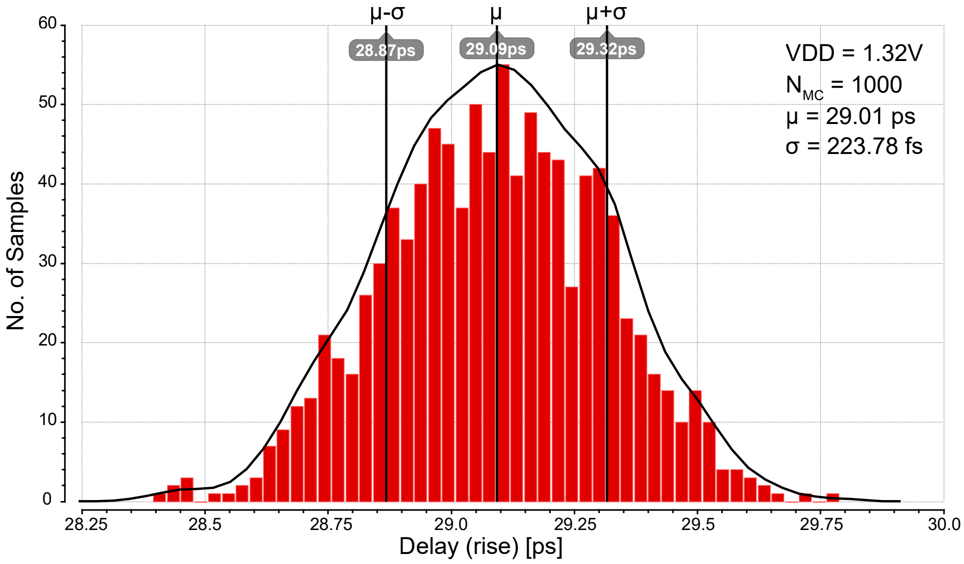


Figure 3.8: Post-layout Monte Carlo simulation of delay element. VDD=1.32

3.2.2 Initial delays

The initial delays are used to create the offset between the interleaved lines. Hence, it is not their absolute delay that is important, but rather their relative delay to each other. Considering the targeted offset of $10ps$, mismatch will have a significantly larger influence on the result. Therefore it will be necessary to implement a technique to compensate for this. The substrate dopant concentration N_A influences V_{th}. N_A is considered a major source of mismatch due to its statistical total-number variation between devices. However, N_A depends on the source-bulk voltage and is therefore not a finite number [27]. By applying forward bias voltages to the body of NMOS and PMOS devices, the threshold voltage is reduced, which increases the on-current. Applying reverse bias voltages will raise the threshold voltage, thereby slowing down the transistor and reducing the subthreshold currents (static power reduced) [28]. The back-gate effect, also known as body effect, describes this change in threshold voltage as a function of the change in bulk-source (in this case for an NMOS) voltage with the following equation:

$$V_{th_N} = V_{th_0} + \gamma\left(\sqrt{|V_{SB} + 2\phi_F|} + \sqrt{|2\phi_F|}\right) \tag{3.3}$$

$$\gamma = (t_{ox}/\epsilon_{ox})\sqrt{2q\epsilon_{Si}N_A} \tag{3.4}$$

where V_{th_N} is the resulting threshold voltage with an applied substrate bias V_{SB}. V_{th_0} is the threshold voltage at $V_{SB} = 0V$, ϕ_F is the fermi potential of the body, the body effect parameter γ depends on oxide thickness t_{ox}, the permittivity of the oxide ϵ_{ox} and silicon ϵ_{Si}, and the doping concentration N_A. q is the elementary charge.

Note that with a statistically inhomogeneous spatial distribution of the dopant atoms in the channel, there is an unstable back-gate dependence of V_{th} [27], i.e., back-gate tuning modulates the mismatch behavior. With a shared back-gate voltage this could result in a degradation of the V_{th} matching between the transistors. However, since the back-gate tuning is applied separately, each delay element can be calibrated independently. The NMOS in the first inverter of each delay element is implemented with a separate back-gate voltage. Biasing V_{B_N} will, within limits, allow us to tune its delay - which we can use to compensate for any deviance from its expected value or its relative delay to the two other. The expected variation revealed by the MC simulations (Table 3.4) indicate that tuning only the first stage should be sufficient. The tuning range could be extended by implementing additional tuning on the PMOS in the second inverter. However, with an adequate tuning range provided by the NMOS, this seems unnecessary, and would only further extend the number of required voltage sources and complicate the tuning process.

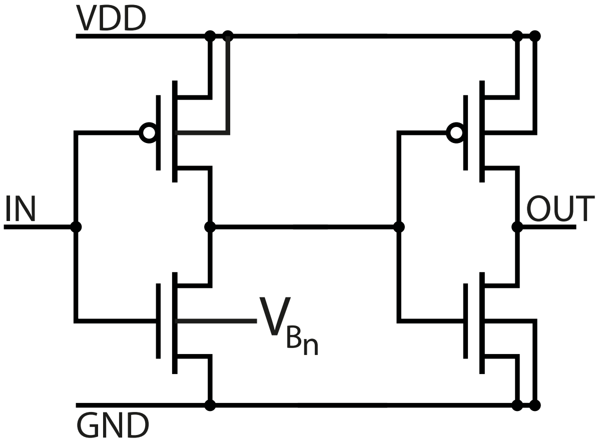


Figure 3.9: Tunable initial delay

Just like the standard delay elements the initial delays are designed unbalanced, emphasizing rising edge performance. In order to ensure a sharp and similar exiting transition the second inverter is designed strong and is identical for all three initial delays (see Table 3.2 and 3.3). The nominal rising edge delay of the elements is adjusted by primarily modifying the NMOS dimensions of the first inverter.

	Row 1		**Row 2**		**Row 3**	
	NMOS	*PMOS*	*NMOS*	*PMOS*	*NMOS*	*PMOS*
n	4	4	4	4	2	4
W	$660nm$	$810nm$	$410nm$	$750nm$	$600nm$	$550nm$
W_{tot}	$2.64\mu m$	$3.24\mu m$	$1.64\mu m$	$3\mu m$	$1.2\mu m$	$2.2\mu m$
L	$120nm$	$100nm$	$130nm$	$100nm$	$150nm$	$100nm$
$Area$	$0.316\mu m^2$	$0.324\mu m^2$	$0.213\mu m^2$	$0.3\mu m^2$	$0.18\mu m^2$	$0.22\mu m^2$
W/L	22	32.4	12.6	30	8	22
W_p/W_n	1.22		1.83		1.83	

Table 3.2: Dimensions for the first inverter in the three initial delays.

	Inverter 2	
	NMOS	*PMOS*
n	6	6
W_n	$630nm$	$1.51\mu m$
W_{tot}	$3.78\mu m$	$9.06\mu m$
L	$100nm$	$100nm$
W/L	38	90
$\frac{Wp}{Wn}$	2.4	

Table 3.3: Second inverter dimensions. The trigger signal enters the delay lines through a strong inverter, providing a sharp transition.

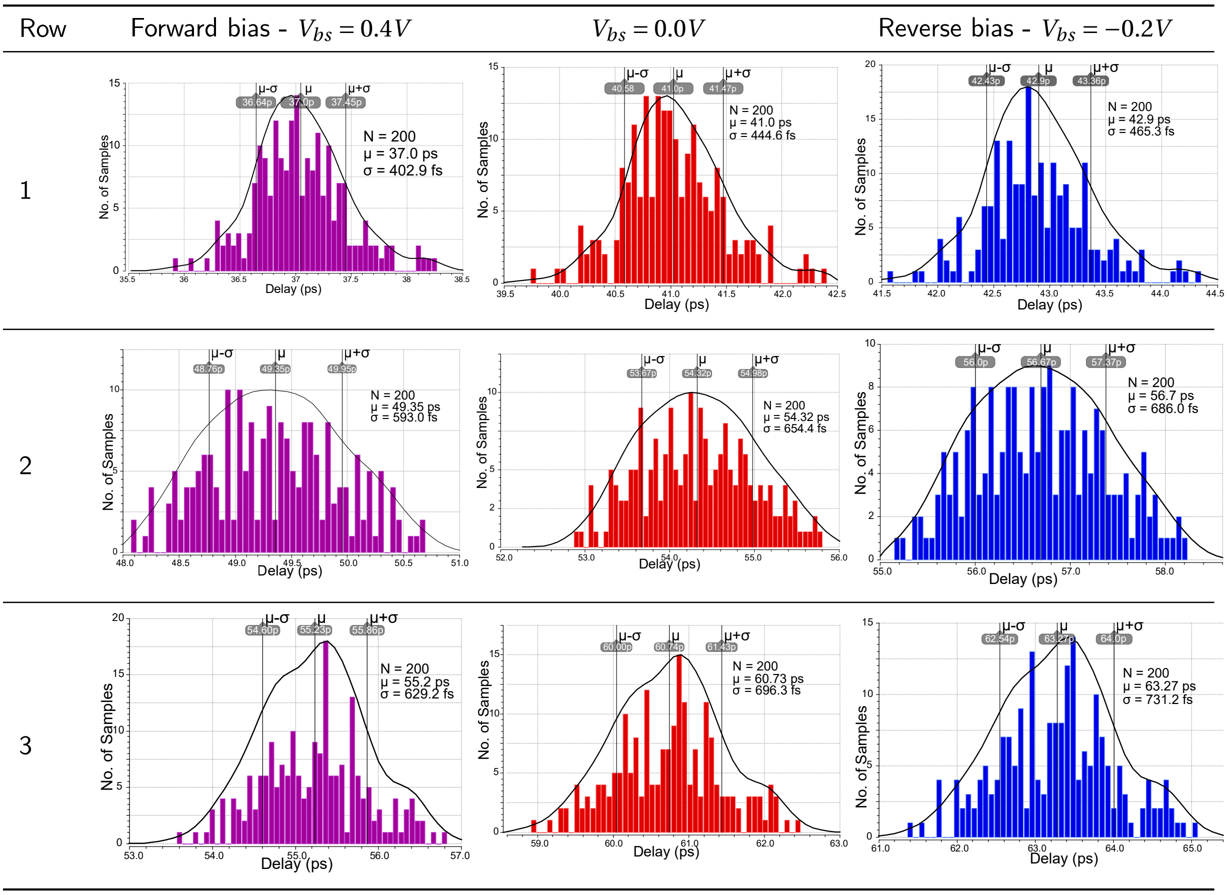


Figure 3.10: Post-layout Monte Carlo simulations of the three initial delay elements with forward biased substrate, nominal substrate and reverse biased substrate. The results are for VDD=1.2V. Higher supply voltages have not been simulated due to the long simulation times, but like for the standard delay elements, a higher VDD is expected to decrease propagation time and reduce the impact from process variations.

Row	μ [ps]			σ [fs]		
	$V_{bs} = 0.4V$	$V_{bs} = 0.0V$	$V_{bs} = -0.2V$	$V_{bs} = 0.4V$	$V_{bs} = 0.0V$	$V_{bs} = -0.2V$
1	37.0	41.0	42.9	403	445	465
2	49.4	54.3	56.7	593	654	686
3	55.2	60.7	63.3	629	696	731

$A \downarrow \sigma \uparrow$

$V_{bs} \uparrow \sigma \downarrow$

Table 3.4: Extracted data from MC simulations

The relative tuning ranges for the three delay elements are approximately the same for both forward ($\Delta\mu \approx -9.25\%$) and reverse ($\Delta\mu \approx +4.44\%$) biased substrate. When forward biasing the back-gate we can see a consistent improvement with regards to mismatch across each element. This is expected since a forward biased transistor will have a reduced V_{th}. Again, we witness that increasing V_{ov} has a positive effect on the influences of mismatch. Also, we observe the inversely proportional relationship between the transistor area and mismatch ($\delta V_{th} \propto 1/A$). As the transistors dimensions for the slower delay elements are reduced, we see a direct impact on mismatch behavior across all substrate biases.

3.3 Samplers

The samplers are placed following each delay element, where they tap the propagating trigger pulse and sample the input signal. These samplers share the same input, namely the thresholded continuous-time input signal. When sampling is initiated the trigger pulse travels down the delay lines, triggers the corresponding samplers in turn, which sample the digital input signal at that specific time. The 1-bit samplers are rising-edge triggered master-slave D flip-flops, which are composed of two gated D latches in series, with complementary clock phases (by "clock" it is here referred to the CTBV trigger pulse). The clock signal is distributed internally to minimize load on the delay elements. The D flip-flops are also used in the counters and shift registers. Their operation is descibed below:

1. clk=0: Master is transparent (passes input value through) while the slave is latched (output of slave is "locked").
2. clk 0→1: Master latches and slave becomes transparent.
3. clk 1→0: Slave latches (output is held at last value seen by master at previous rising edge), master becomes transparent once more.

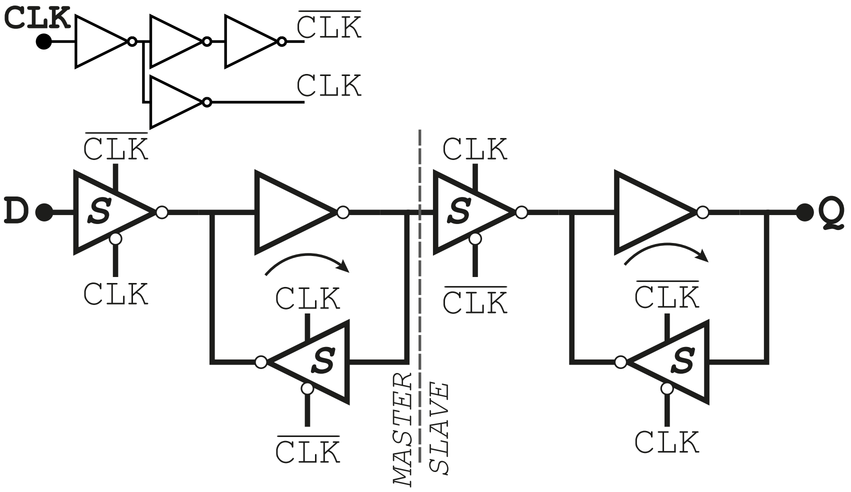


Figure 3.11: Rising-edge triggered master-slave D flip-flop

3.3.1 Timing considerations & metastability

A flip-flop can reach a metastable state if two inputs, e.g., data and clock, are changing at *about* the same time. If the input signal is at an intermediate state between the two defined logic levels when the rising clock edge triggers the flip-flop, it *may* resolve to either 1 or 0, or reach an undefined metastable state between the two levels. To avoid metastability, the input to a flip-flop must be stable in a period around the rising edge of the clock. This period is often referred to as the *aperture*, which is the sum of setup and hold time[2]. Since the flip-flop samples the continuous-time output from the thresholder that can change at any time, there is no way to guarantee that the setup and hold criteria are met. However, clean sharp clock transitions can help to reduce the aperture of the flip-flop. In the event of metastability, a flip-flop could, in theory, remain in a metastable state indefinitely, but in reality, it won't. Any noise present will nudge the state of the flip-flop causing it to settle from the metastable state into either 0 or 1. Flip-flops are often characterized with a settling time, which is the maximum time they are expected to remain metastable under specified conditions. In an attempt to suppress the probability of metastable states causing glitches in the counters, a delay element with starved inverters has been incorporated as seen in Figure 3.12 (B) to give the flip-flop ample time to settle in the case of metastability. If the D flip-flop samples a "1", this will not increment the counter until the delayed trigger pulse reaches the AND-gate. The starved delay element is detailed in subsection 3.6.1.

[2] Setup and hold time is the minimum time the input should be held steady before and after the clock event to ensure that the flip-flop samples the data correctly

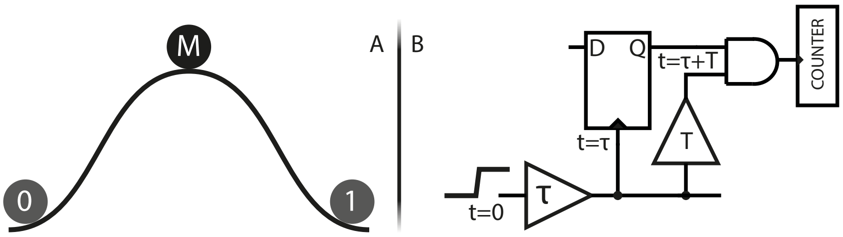


Figure 3.12: **A:** Metastability can be compared to a ball on a hill peak, where it is in a "stable" state at the top. The slightest breeze would cause the ball to roll down to one side or the other. **B:** A starved inverter delay element, $T \approx 0.5ns$, is used to provide the sampler adequate time to settle and avoid glitches on the counter.

3.4 Counters

The counters are implemented as simple asynchronous ripple counters, where the D flip-flops are configured to toggle their state on the clock edge by using the complemented output ($\overline{Q}$) as feedback to the input, equivalent to a T flip-flop. The first flip-flop is clocked by the output of the sampler, while the clock input of the subsequent flip-flops are tapped from $\overline{Q}$ of the preceding flip-flop. This way the clock signal appears to ripple its way from the first to the last stage as the counter increments. The counters are 16-bit, which means that a maximum of $2^{16} - 1 = 65.535$ repeated measurements can be performed without running the risk of overflowing the counters. This design choice is primarily due to the limited area on the chip; the number of bits can easily be increased to accomodate higher processing gain, at the expense of increased area.

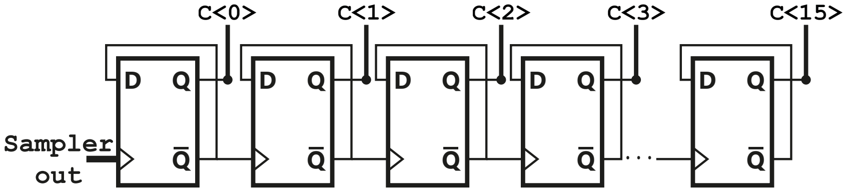


Figure 3.13: Schematic of asynchronous ripple counter

3.5 Input/Output

3.5.1 SPI module

All data are transfers with the chip goes through the Serial Peripheral Interface (SPI). SPI is a synchronous serial bus between one master and one or more slaves that require only four signals. This makes it ideal for embedded systems such as this. The on-chip SPI module used in this thesis was developed by Håkon Hjortland [5]. The software, originally developed by Hjortland for a single board microcrontroller, and further developed by Øystein Bjørndal [10] for Raspberry Pi, were used as a starting-point and ported over to be compatible with Beaglebone Black. A 64-bit (only 14 used) control bit register, supplied by Øystein Bjørndal, is used to set various control signal bits in the sampler and quantizer over SPI. The various control signals are explained in Table 3.5.

Control bit(s)	Purpose
GAIN<2:0>	Quantizer threshold gain control
TRIG-ENABLE	Enable/Disable TRIGGER input
SEL-DATA	Sets 2:1 MUX selecting sampler input from pad or Quantizer
DELAY<3:0>	Sets programmable delay
MODE	Sets MODE (read/shift) in shiftregisters
REG0	A single bit is output from register to pad to test if SPI works
nEN<2:0>	Enable/Disable output from initial delays to pad for characterization

Table 3.5: Control bits set over SPI to control the operation of the system

3.5.2 Shiftregisters

The sampler data is to be read out over SPI as a bitstream. The sampler data is converted to a bitstream using a daisy-chain of 16-bit Parallel-in/Serial-out Shift Registers (PISO) connected to the counters. In read mode, the PISO registers load the counter data in a parallel format, latching all the bits from their corresponding counter simultaneously (one clock pulse). As the PISO register enters shift mode, the data can be read out one bit at a time in a serial format. The mode selection bit is set in the control bit register, and the two clocks are output by the on-chip SPI module.

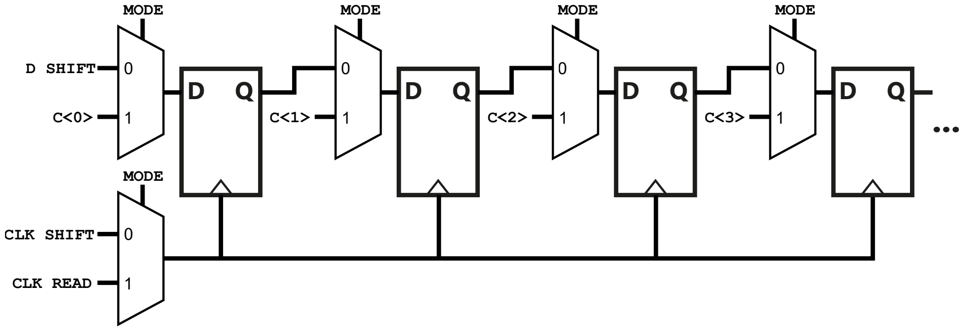


Figure 3.14: Parallel-in/Serial-out Shift Register

3.6 Misc. supporting circuitry

3.6.1 Delay element with current starved inverter

By shorting gate and drain of a CMOS transistor it becomes diode-connected. In the configuration shown in Figure 3.15 the diode-connected transistors act as two-terminal resistors, restricting the charging/discharging current of the first inverter, i.e., extending the propagation delay. This is because the diode-connected transistors are always in saturation ($V_{DS} = V_{GS} \rightarrow V_{DS} \geq V_{GS} - V_{TH}$), where $r_{DS} \approx 1/g_m$ [29]. The current starved delay elements do not have to be very accurate, as long as they delay sufficiently for the D flip-flops to settle in case of metastability. A delay of ~ $0.5ns$ has been considered sufficient.

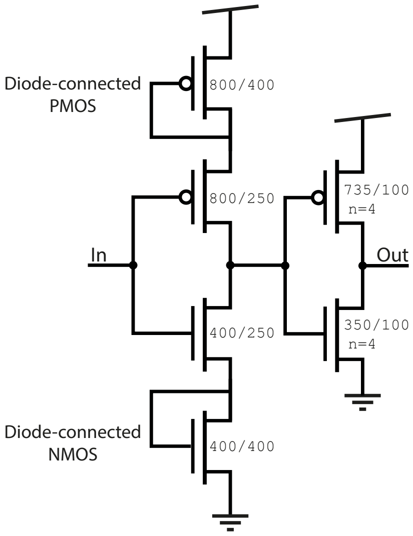


Figure 3.15: Delay element with starved inverter

3.6.2 Programmable delay element

A simple SPI-programmable delay element has been implemented at the trigger input. Its purpose is to allow some control over the time frame to be sampled by delaying the initiation of the sampling sequence. It is controlled over SPI and consists of a cascade of starved inverter delays, providing only coarse adjustment across 8 settings; with the first being only the inherent delay of the 2:1 MUX ($\sim 200ps$), and increments in steps of $\sim 500ps$ up to $\sim 3.7ns$.

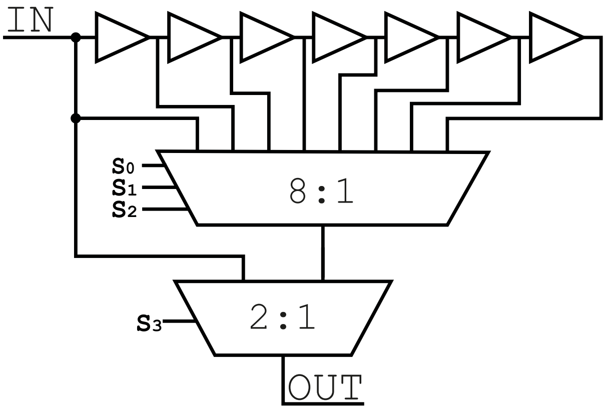


Figure 3.16: SPI-programmable delay element

3.6.3 Schmitt Triggers

Following each of the analog input pads on the chip are Schmitt Triggers. A Schmitt Trigger provides hysteresis with two threshold voltage levels, where it retains its level until the input changes sufficiently to trigger a change. With an anlog input signal it acts as a fixed threshold comparator, converting the analog input signal to a digital output signal. It is also useful with a digital input, since its hysteresis can avoid bit errors in noisy signals. The Schmitt trigger has proven invaluable during measurements by allowing us to bypass the LIDAR system and characterize the system with a known signal.

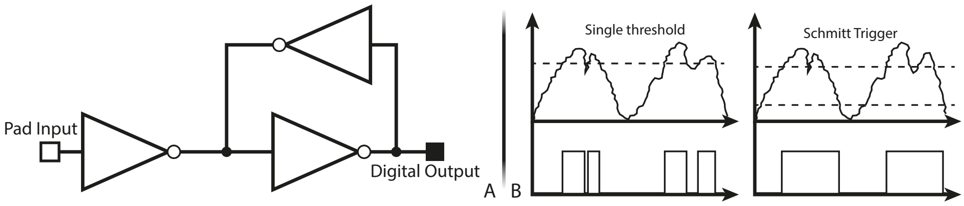


Figure 3.17: **A:** Schmitt trigger schematic. **B:** Schmitt trigger vs. regular single-threshold buffer.

3.6.4 Fan-out buffering

For signals with high fan-out, adequate buffering is added to provide clean and fast transitions. These signals include the thresholded input signal, the SPI clock signals, as well as the mode select signal for the shift registers and the asynchronous reset signal. Also, all outputs to pad are buffered significantly due to the high capacitive load of the analog pads.

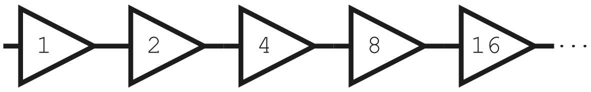


Figure 3.18: Cascaded buffers of increasing size are used to increase the fan-out.

3.7 Complete system schematics

Figure 3.19 portrays how the different components come together to make the proposed sampling system. The figure shows the schematic representation of the first sampling line. The externally applied signal SHIFT IN is the value (0 or 1) that will be shifted in during readout. Figure 3.20 provides a system overview showing the sampling system as a whole.

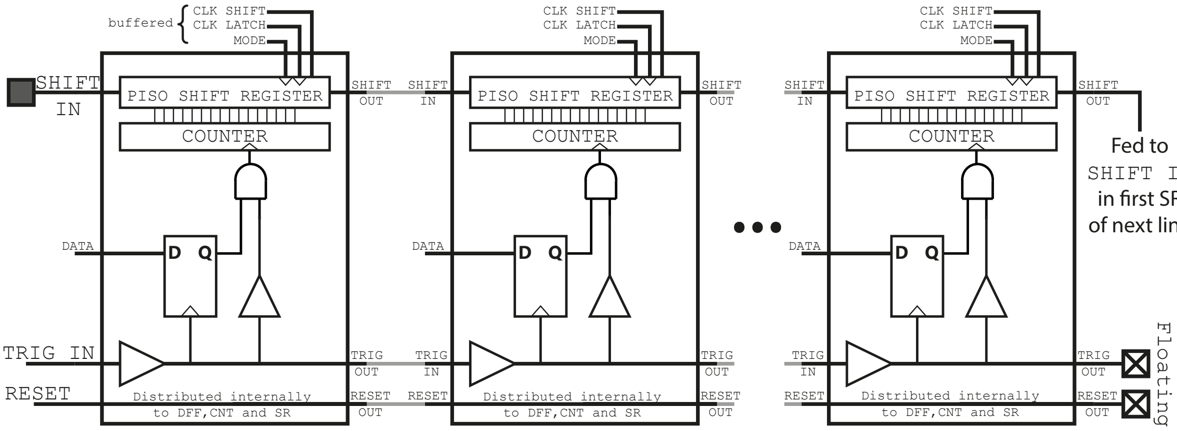


Figure 3.19: Sampler line

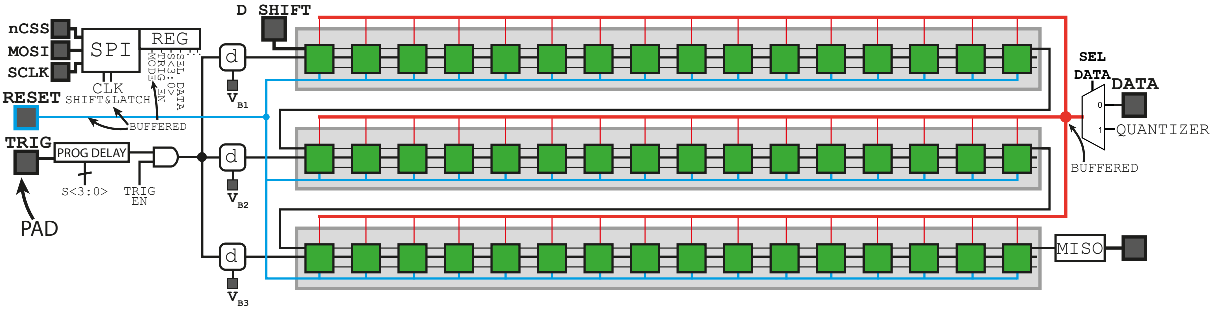


Figure 3.20: Complete system block diagram that shows the sampling system in the package. Note that Schmitt triggers and buffers are not depicted.

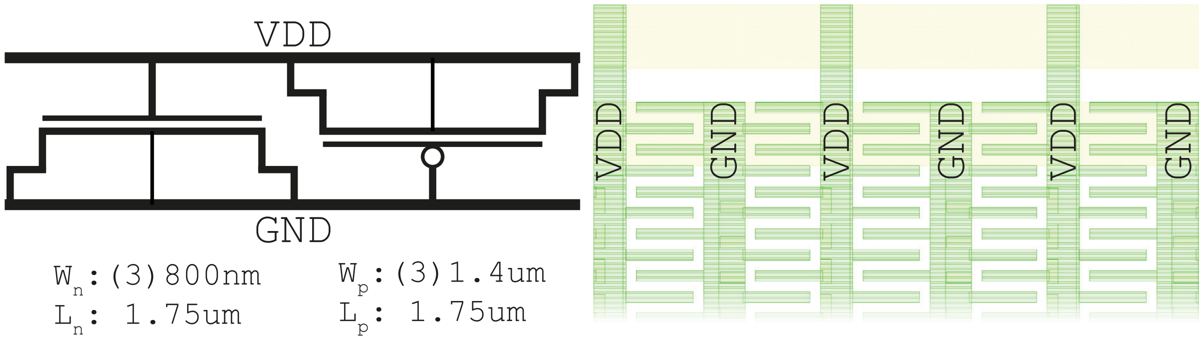
3.8 Power Supply

As stated in section 2.7, power-supply noise is directly correlated to the switching activity and current consumption of the circuits on the grid. For this reason, the system has been implemented with two separate power grids to reduce the impact of these variations on the delay lines. One (VDD) drives the majority of the system, while the second (VDDA) is solely for the delay lines. This way the power-hungry blocks of the system will not directly influence the delay lines through power-supply noise caused by the accumulated switching currents of their combined switching. Furthermore, this allows us to apply gate delay tuning on the delay lines by regulating VDDA without affecting the non-critical system blocks.

3.8.1 Decoupling

In order to suppress noise on the power-supplies ample on-chip decoupling (or bypass) has been implemented between the supply rails. The decoupling capacitors hold an electrical charge that is released to the power grid whenever a transient spike occurs. This helps help to minimize the supply-voltage noise generated by the switching devices by providing a low-impedance supply (fast) which decouple the power supply from the switching circuit. The decoupling capacitors are implemented as MOSCAPS since they have been shown to be the most area-efficient due to the thin dielectric layer provided by the gate-oxide thickness [30]. Banks of parallel decoupling capacitors have been placed close to the switching circuits where there has been room (see Appendix D). Sufficient decoupling has been particularly important for the "digital" supply (VDD) to decouple the samplers, counters and shift registers. Furthermore, even though the delay lines are on a separate supply, decoupling has also been implemented here. Even though the dynamic power consumption will be minimal, it is only sensible to provide decoupling here as well considering the critical nature of these lines.

By distributing the power nets in an interdigitated structure with minimum distance as shown in Figure 3.21b, the lateral capacitive coupling between two traces has been taken advantage of to introduce some decoupling in the layout of the power nets as well.



(a) CMOS decoupling capacitor consisting of both NMOS and PMOS capacitor

(b) With alternating lines of VDD and GND in an interdigitated pattern we increase the capacitive coupling.

Figure 3.21: On-chip decoupling strategies

Chapter 4

Measurements & Characterization

In this chapter, the completed sampling system is placed in a testing environment where it is characterized and evaluated. All measurements were performed in-house in an ESD secure lab.

4.1 Measurement setup

In the measurement and evaluation of the sampling system, an externally generated RF input signal was used to emulate the pulsed LIDAR input signal. The RF signal generator (R&S SMF 100A) is capable of excellent phase offset adjustments down to increments of $0.1°$, which corresponds to a temporal resolution equal to $\sim 278fs$ at $1GHz$. We have taken advantage of this fine phase accuracy to characterize our sampler. The RF signal generator did not use its own reference oscillator; instead, it was provided with the reference clock of the clock generator. This way we ensured that the two instruments remain synchronized during testing, which allows us to perform repeated measurements without randomizing the input between measurements; a requisite for coherent signal reconstruction. In the absence of a working LIDAR system, the quantizer became unavailable to us. Fortunately, the embedded Schmitt trigger at the RF signal input doubles as a fixed threshold comparator, providing sharp transitioning square waves from the RF sine input. All measurements were performed with the applied supply voltages of $VDD = 1.2V$ and $VDDA = 1.3V$ (Delay lines).

The clock generator, CG635 from Stanford Research Systems, generates very stable clock signals up to ~2 GHz, with a jitter resolution of $< 1ps$ RMS and fast transition times at $< 100\ ps$ (20% to 80%). The phase offset adjustment is limited to $\sim 14ps$, which is why any phase adjustment has been made on the more accurate RF generator instead. CG635 uses a high-performance internal rubidium crystal timebase with a stability of $0.5ppb$. This high accuracy oscillator substantially reduces the low-frequency phase noise on the output and provides a highly accurate synchronization between the clock and RF source by using it as a shared reference timebase between the two instruments. In addition to handling the chip I/O communication over SPI, the Beaglebone Black [31] were also controlling all the instruments through the SCPI interface standard, either over TCP/IP or GPIB. This way more elaborate measurements could, to a certain extent, be automated and the data collected and processed in post. One drawback with this clock generator has been the inability to send accurate bursts of pulses. Unlike some clock generators, CG635 is not implemented with the option to pulse a specified number of times. The solution has therefore been to enable the output, then disable it after a specified amount of time controlled by the Beaglebone Black. Due to a notable latency in all communication between Beaglebone Black and CG635 over the GPIB interface and inaccuracy in the BBB timing, the clock frequency was kept at $1MHz$ to avoid significant variation in the number of pulses generated between iterations. Still, there was some variation in the number of pulses (±5000) generated between iterations. Since the sampling results are normalized in further post-processing, this was not of any concern during our characterization of the system as long as overflow of the counters is avoided. However, in a swept threshold sampling scenario, CG635 would appear too uncontrollable in this current configuration. In Figure 4.1 below the measurement setup is illustrated.

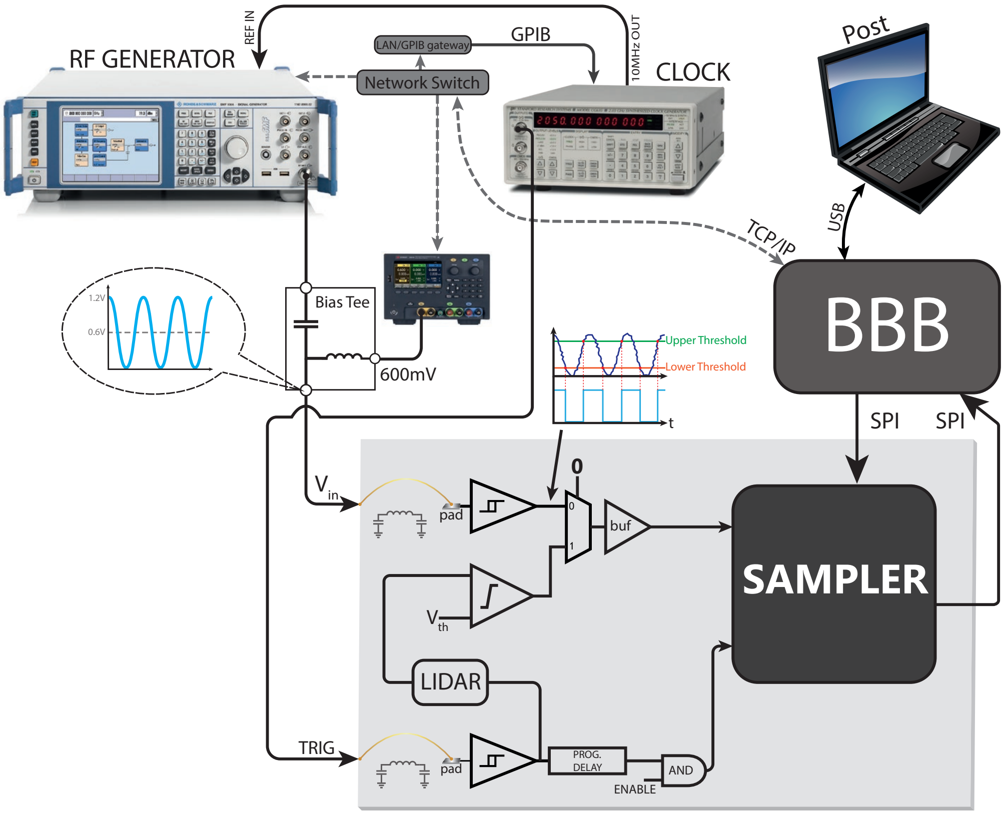


Figure 4.1: Measurement setup

4.2 Characterization of the delay lines

To assess the propagation delay of each element in the system a simple, yet effective technique has been used. The input signal is phase shifted to the starting position (ϕ_0) where the rising edge of the input signal occurs just outside of the samplers frame. By performing a full sampling at incrementally increased phase offsets, with steps of $\Delta\phi = 0.1°$ from ϕ_0 to $\phi_0 + 220°$, the sampled results of the counters will catch the movement of the transition along the delay lines as illustrated in Figure 4.2. This way a full mapping of the delay line characteristics is captured and can be analyzed with a Python script that identifies the phase shift required to move the rising edge (50%) from one sample to the next and relating that phase shift to corresponding time using the relationship in Equation 4.1. Each line was studied independently across three repeated data collections in order to ensure on-chip sampled measurement accuracy. The raw data from one such data collection has been included in Appendix C.

$$\tau_n = \frac{\Delta\phi}{360° \cdot f} = \frac{\phi_n - \phi_{n-1}}{360° \cdot 1GHz} \tag{4.1}$$

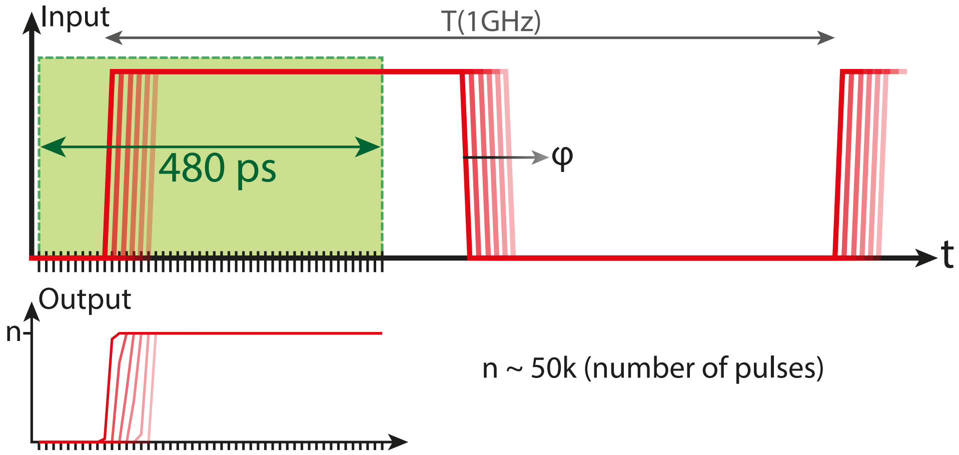
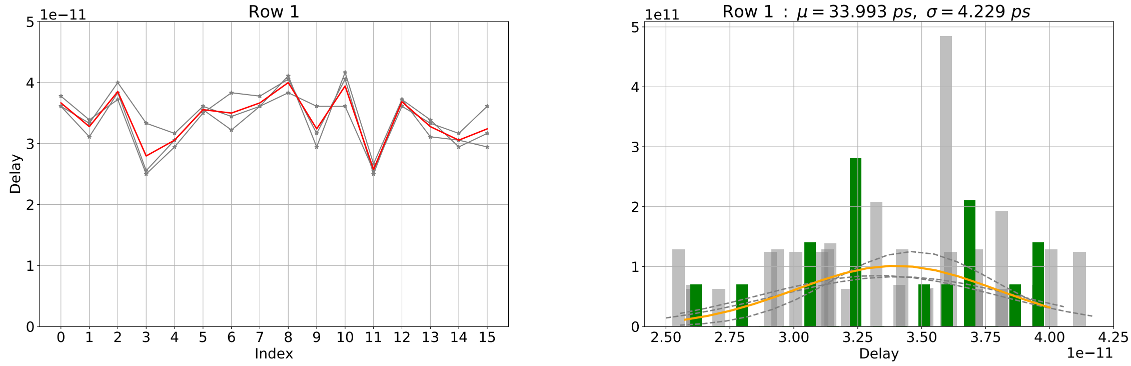


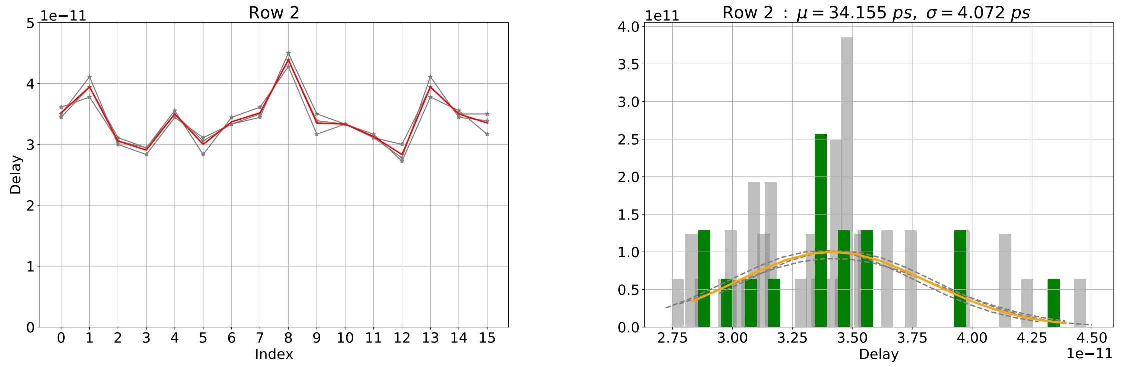
Figure 4.2: Delay line assessment is achieved using accurate phase shifting of a sharp transition.



(a) Delays

(b) Histogram distribution, with calculated μ and σ

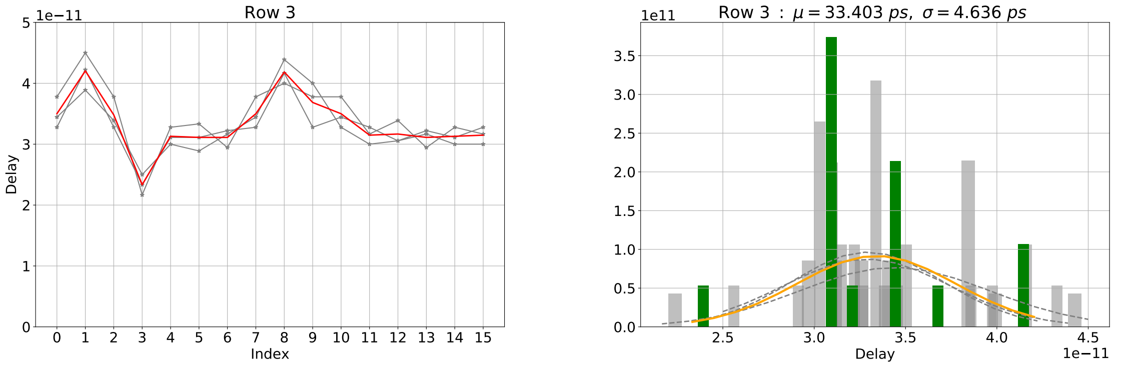
Figure 4.3: Measurement results for delay line 1



(a) Delays

(b) Histogram distribution, with calculated μ and σ

Figure 4.4: Measurement results for delay line 2



(a) Delays

(b) Histogram distribution, with calculated μ and σ

Figure 4.5: Measurement results for delay line 3

The measured unit delays for each individual lines are presented in Figure 4.3-4.5. The red traces are the average of the three repeated measurements, while the green traces represent the resulting histogram. The results reveal that the three lines are quite similar with respects to μ and σ. The three lines exhibit a similar response, with an average unit delay $\approx 34ps$, somewhat higher than expected and a spread of more than $4ps$. This can most likely be attributed to additional capacitive parasitics appearing in the fabricated chips, not captured during the post-layout simulations. Also, the standard deviation is larger than expected possibly indicating even larger devices are required. Since the modeled delay elements appear a tad too weak to produce the required delay unit delay of $30ps$, the equivalent sampling rate will be reduced somewhat from the targeted sampling rate of $100GS/s$. Furthermore, the high degree of variation can prove degrading to the sequentiality of the sampler.

4.3 Sequentiality

A vital criterion for the integrity of the sampler is the requirement for a sequential sampling order in the interleaved system, i.e., *sequentiality*. The order in which the samplers are triggered is determined by the propagation of the rising edge trigger through the delay lines and in particular the relative offset between the initial delays. This order is predefined in the system design, but as the results above indicate there is a high degree of variation in the lines. It is plausible to expect that, at least in an untuned state, individual samplers might "cross over" their neighbor causing an incorrect sampling sequence, as illustrated in Figure 4.6. Initial measurements confirm this to be the case. By adjusting the input signal phase offset between repeated measurements, we can evaluate the triggering order by studying the sampled output. While untuned, we detect several cases of non-sequential behavior in the sampler. One such result is shown in Figure 4.7, where the sampled waveform shows that sampler 12, which taps on line 1, samples before sampler 11, tapped from line 3. In an untuned state, this behavior is not unexpected, after all, it has been the main reason for implementing back-gate tuning.

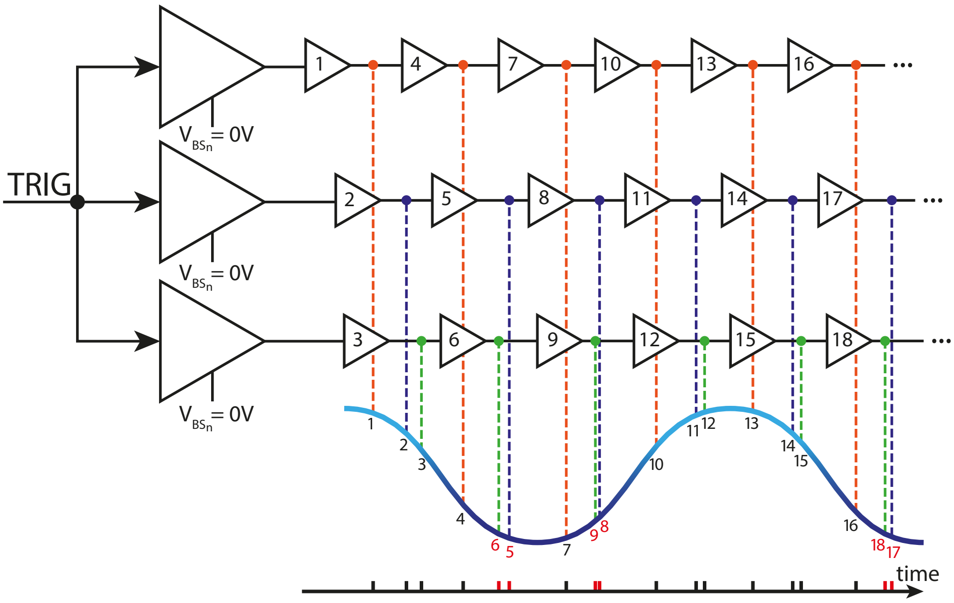


Figure 4.6: Illustration: nonsequential sampling occurs without back-gate compensation

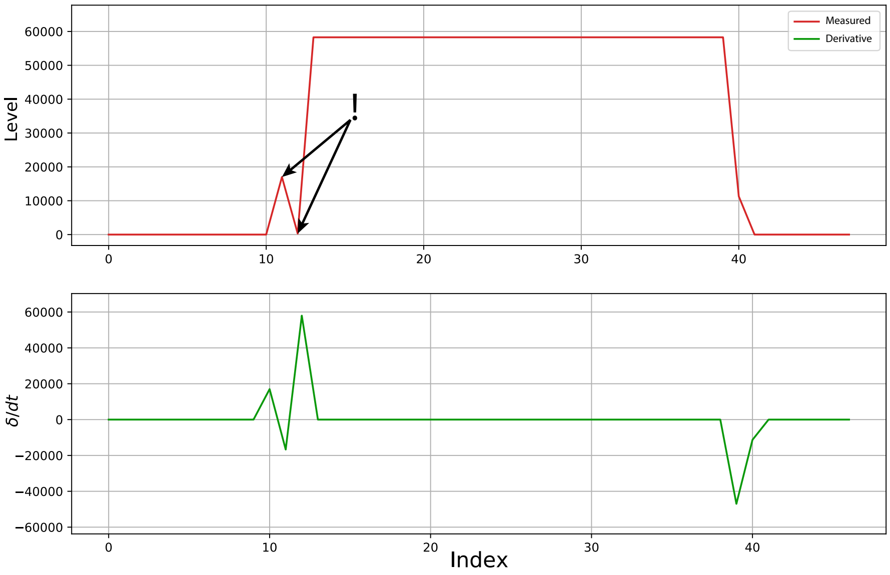


Figure 4.7: Measurement reveals nonsequential behavior in the sampler when left untuned. Sampler 12 is triggered before sampler 11.

At this stage, the primary concern was to attempt to eliminate this behavior. When incorrect sampling sequence was observed the back-gate voltages were adjusted to realign the relative time offset of the delay lines, reasserting sequential behavior at that point. This manual tuning proved time-consuming but eventually allowed us to phase shift the input signal flank through the interleaved system in fully sequential order. Although this initial back-gate compensation has presented promising results, it appears that it will be difficult to compensate to create an interleaved system with (close to) equidistant sampling. In an interleaved sampling system of perfectly equidistant sampling, an index shift should occur with an input phase shift of exactly $\Delta\phi = 3.6°$, which corresponds to the targeted $10ps$ temporal shift. In reality, this is rarely the case. The measurements reveal that the phase shift required ranges from 2° to almost 8°, portraying a reality more as illustrated in Figure 4.8. The high degree of mismatch in the individual delay lines causes a nonuniform delay distribution in the individual lines. Some degradation of equidistant sampling were predicted in our post-layout simulations. However, with a spread of more than $4ps$ the variation is so extensive that achieving an ideal equidistant sampling seems unlikely in its current topology.

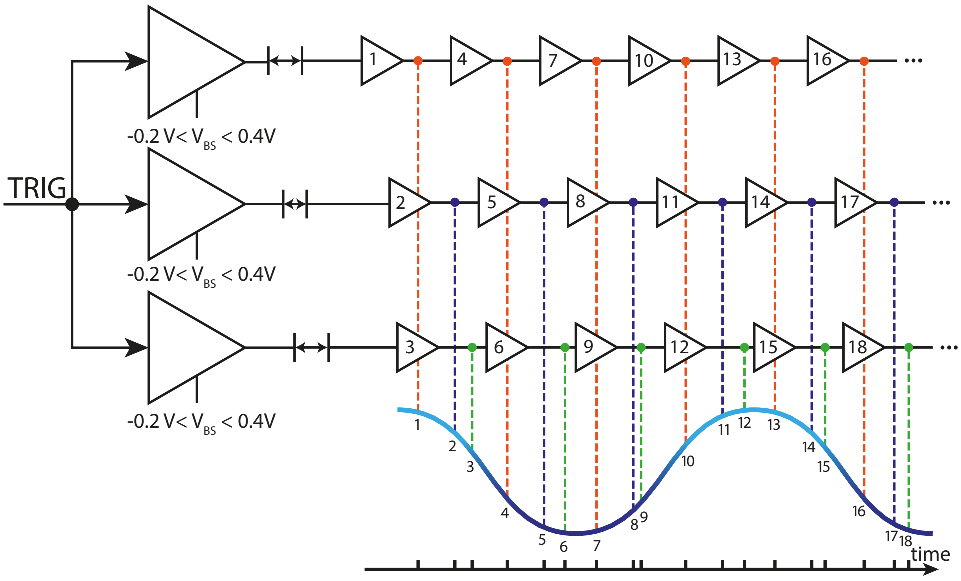
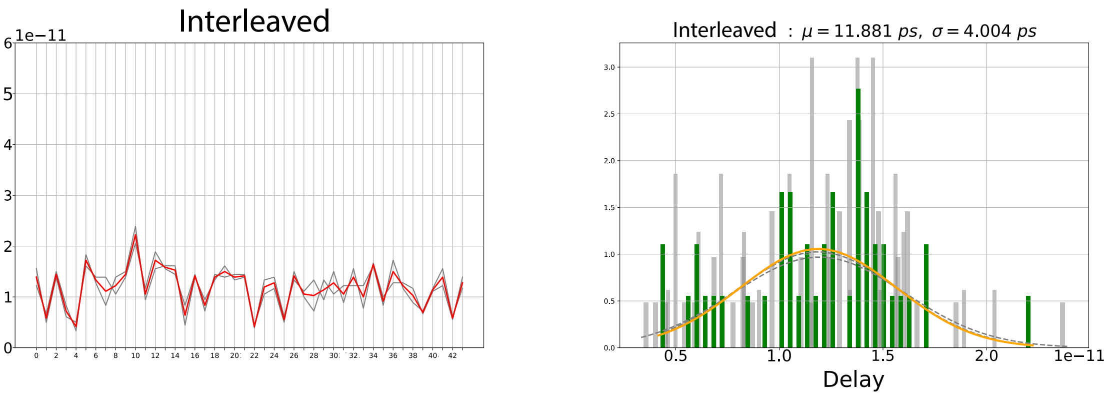


Figure 4.8: Back-gate compensation ensures sequential, but not equidistant sampling.

4.4 Characterization of the interleaved system

During the design phase, additional outputs to pads were integrated at the outputs of the initial delays for testing and debugging. However, these signals required significant buffering before pad. With a notable level of mismatch in these buffers, the outputs cannot be considered reliant for alignment purposes of the initial delays. Furthermore, with substantial delay variation between the SMA cables used to transmit the signals to an oscilloscope, no relevant information can be collected from this, apart from confirming that the trigger pulse(s) in fact enter the delay lines. Significant effort was invested in manual back-gate tuning in an attempt to improve the alignment of the delay lines further. This task proved very challenging and did not yield meaningful improvements to the results. An optimal tuning at an early stage of the interleaved system caused non-sequential behavior further out and vice versa. Therefore, a compromise was necessary in order to ensure that the sampler maintained a correct sampling sequence. With the back-gate compensation in place, the phase shifting measurements were repeated; this time to evaluate the propagation of the trigger pulse through the interleaved sampling system.



(a) Delays

(b) Histogram distribution, with calculated μ and σ

Figure 4.9: Measurement results of the interleaved delay lines. Note how it has only computed to index 42 of 47. This were later discovered to be caused by a bug in the script which evaluated the data.

Recognize how the delay between triggered samplers varies widely, with outliers of shortest delay being only $4ps$ and the longest being $\sim 23ps$. However, the majority are clustered around $\sim 12ps$ and the standard deviation is calculated to be $4ps$. The unit delay spread of $4ps$ is larger than expected and is hard to explain. The instrumentation and test setup were carefully re-evaluated for measurement errors and did not reveal any oversights. Even so, based on the nonuniform propagation in each line, it comes as no surprise that the interleaved sampling exhibits similar characteristics. Based on the calculated mean delay of the system, $\mu = 11.88ps$, the sampling rate can be equivalated to $84GS/s$; albeit it is not equidistant. Note how the zig-zag pattern that appears in the interleaved system confirms our expectations illustrated in Figure 4.8. Following a short delay, there is often a long one and vice versa. Measurements performed on a second chip (not shown) presented similar results of mean delay and spread, but required entirely different back-gate voltages for alignment, indicating a notable influence from inter-die variations as well. During the tests performed, no apparent cases of metastability in the samplers causing the counters to trigger incorrectly were observed. It seems that the added delay element (see subsection 3.3.1) proved adequate in suppressing this issue.

4.5 Power

The highest current consumption during our measurements was $2.7mA$ on VDD and $0.1mA$ on VDDA. This gives a peak power consumption of the sampling system of only $\sim 3.4mW$. This indicates that the proposed CTBV design approach provide almost two orders of magnitude faster sampling rate than a strictly clocked system even with low power consumption. Note that with an externally generated clock signal, and since the quantizer was not used for thresholding, this result only displays the power consumption of the actual sampling system. In a swept threshold sampling, the quantizer is expected to cause a higher power consumption.

$$\mathrm{P}(VDD) = 2.7mA \cdot 1.2V = 3.24mW \tag{4.2}$$

$$\mathrm{P}(VDDA) = 0.1mA \cdot 1.3V = 0.13mW \tag{4.3}$$

Chapter 5

Discussion and Further Work

In this thesis, the primary effort has been on obtaining an understanding of the design requirements for continuous-time chip design using inherent gate delay of standard CMOS. This perspective has enabled us to design the delay elements to the best of our ability while acknowledging how the catch-22 dilemma of mutually conflicting conditions between speed and stability has required compensation using back-gate tuning. The empirical results have to some extent confirmed these predictions, however, the real-world implementation has shown that production variations are in fact even more severe than predicted by modeling. In our efforts to maximize the speed of the delay lines the inherent process variation has a dominating impact on the performance. This indicates that the models used to perform the design analysis for optimization has been too inaccurate to predict the implemented characteristics of the delay lines adequately. The sampling system on the finished chip has been evaluated to produce a sequential but not equidistant sampling at a rate of $84GS/s$. The reduced sampling rate is attributed mainly to the unit delays being too slow, producing an average delay of $34ps$. Furthermore, the substantial degree of variation in the lines is indicating that having only initial back-gate calibration is insufficient for ensuring an equidistant sampling. The back-gate tuning allowed us to avoid nonsequential sampling but did not extend so far as to provide a cure for the variation in the lines. Our results show that we have indeed pushed the boundaries of the technology, and the effects of this are apparent. Even so, what we have accomplished in our continuous-time sampling system show great potential for further improvements - the two immediate aspects worthy of attention being gate delay and variation/compensation.

Although the sampling appears nonuniform, this characteristic also appears systematic since the measured delays show little to no deviation between repeated measurements. The Shannon sampling theorem for nonuniform sampling states that a band-limited signal can be entirely reconstructed from its samples as long as the average sampling rate satisfies the Nyquist condition [32]. This means that we may reconstruct signals up to $42GHz$ by using somewhat advanced reconstruction algorithms such as Lagrange interpolation in post.

We were not able to reach the targeted unit delay of $30ps$ in our architecture. An immediate improvement can be gained by decreasing the load on the delay elements. While the D flip-flop and subsequent delay element are unavoidable loads on the delay element, the starved delay element which in this design taps the trigger pulse directly from the line introduces an additional capacitive load that could have been avoided with a different solution. Since the starved delay is not a vital part in the timing of the sampler, a more viable placement would have been to incorporate it in the D flip-flop instead as depicted in Figure 5.1. This would, of course, require careful sizing of the clock distribution inverters in the D flip-flop to compensate for the added load. Furthermore, there are arguments to be made of using a more conservative technology model. This could potentially have foreseen the delay inadequacy during the design phase and commanded even larger devices. A more global improvement can be found in transitioning to faster technologies. Since the gate delay is reduced for finer pitch technologies, the sampling rate would directly benefit from a more high-end process than 90nm TSMC CMOS, such as 65nm or even 55nm low-power CMOS.

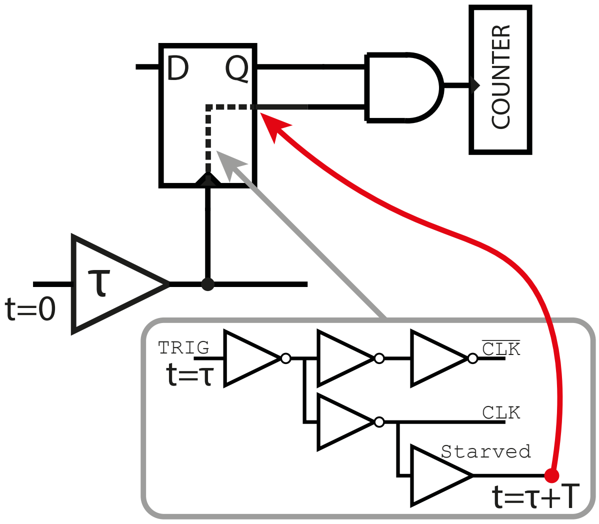


Figure 5.1: The unit delay can be decreased by incorporating the starved delay element into the D flip-flop.

The high spread in the delay lines made the interleaved system far from equidistant and even nonsequential prior to tuning. The back-gate tuning proved to be an arduous and somewhat ineffective operation, which primarily allowed the sampling to return to sequential order, yet did little to improve the uniformity of the sampling. The tuning also differed widely between the two chips tested, signifying a notable contribution from inter-chip variations as well. These findings indicate that extended and potentially automated calibration routines in the system can have a significant impact on further improving the performance of the sampling system. One way to improve the sampling could be to distribute additional tunable elements at regular intervals throughout the lines. This way tuning can be performed for a more optimal equidistant sampling in each "section" and re-evaluated and compensated when necessary at these intervals. While this should improve the performance, it would also further complicate the complexity of the tuning process and make manual calibration a nightmare.

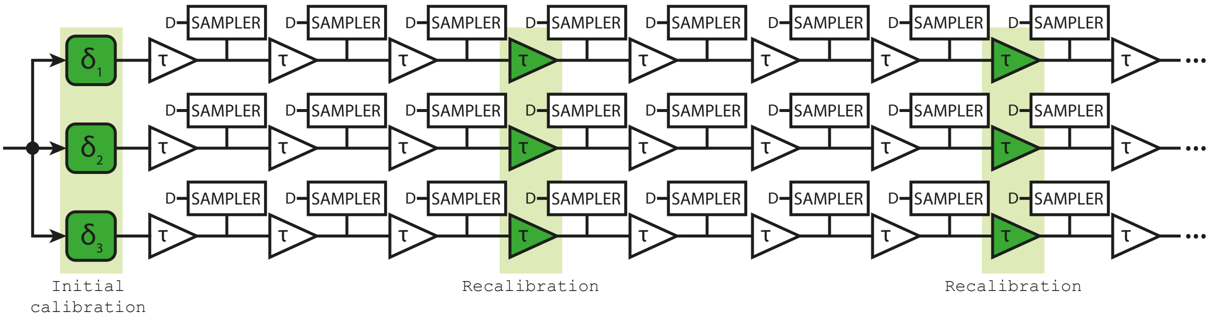


Figure 5.2: Proposed delay line topology with extended back-gate calibration.

With the already challenging calibration routine with just three tunable elements, efforts would be well spent in investigating options for more automated or self-compensating calibration routines, especially when considering the added calibration complexity of extending the number of tunable delay elements. The novel high-precision calibrated delay element presented in [13] has been experimentally shown to provide precise delay calibration down to $1ps$ RMS error by using a ring oscillator-based calibration circuit. This calibration technique or a modification of it could greatly simplify the calibration routine during operation; however would also involve an added complexity to the chip.

The significant presence of device mismatch in the delay lines has resulted in a high tuning requirement in the system. It has been made clear that in the current topology, the tuning routines must be expanded if we are to ensure a sequential *and* equidistant sampling sequence throughout the interleaved system. However, considering that the sampler is intended to assess the properties of blood in veins which are only a few millimeters thick, the only interesting reflections will come from one or maybe two samples inside the vein.

($d = c \cdot \tau = 3mm$). By decreasing the temporal window (shorter lines) the initial tuning will become more potent since the accumulated degradation due to the spread in the lines will be reduced. Implementing a high precision programmable delay element at the input will allow us to move the now smaller window around with fine increments in order to adjust for the targeted depth we are studying. Furthermore, since the sampled signal is coherent we can expand the sampled frame if needed by performing samplings at incremented depths and stitching/combining the results in post.

One way to reduce the depth of the delay lines without affecting the temporal window could be to increase the number of parallel delay lines. Instead of having three interleaved lines, one can have for example five. Assuming that the more elaborate calibration routine of having more lines can be maintained with for instance the self-calibrating initial delays mentioned earlier, each line can be shorter and consist of delay elements with a longer unit delay ($\tau = n \cdot 10ps$, line depth $= \frac{\text{targeted window}}{\tau}$). Although the *relative* variation will be reduced with slower delay elements, the absolute variation and jitter will increase. This topology would only increase the design challenges of mismatch further and besides, also depend more heavily on tuning. For this reason, this topology is not considered a viable strategy for improvement and should be contemplated carefully before attempted.

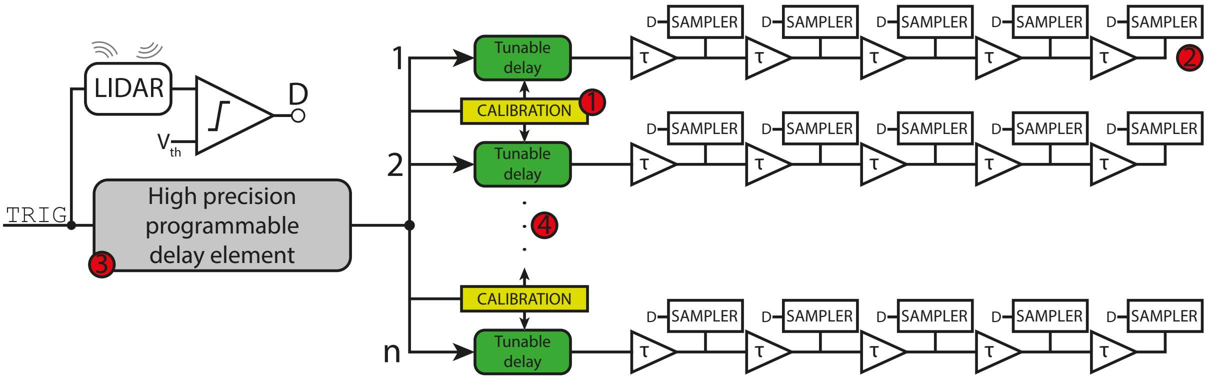


Figure 5.3: Various suggested improvements. **1:** Automated calibration. **2:** Reduced delay line depths **3:** High-precision programmable delay element at the trigger input. **4:** Increased interleaving.

In this thesis, we have presumed the dominant source of variations to be the static production variations. Our results indicate that with further improvements to the compensation strategy, the effects of mismatch can potentially be reduced to a tolerable level. However, as the static effects are subdued, other effects will appear more prominent. Dynamic variations such as jitter, temperature dependencies, and power supply noise can at some stage become the dominant limiting mechanisms of a continuous-time system, and require particular attention.

Chapter 6

Conclusions

In this thesis, we have presented a novel high-speed and low-power sampling solution successfully implemented and measured in TSMC 90nm low-power CMOS. It has with its continuous-time sampling solution achieved a sampling rate close to two orders of magnitude faster than the clocked alternatives. Three interleaved delay lines with initial delay offsets for alignment have been modeled to generate the timing. The added design challenges of process variations have been addressed by implementing back-gate tuning on the initial delays to provide calibration and delay line realignment during operation. We were able to achieve sequential sampling for all 48 points, with an equivalent sampling rate of $84GS/s$. Moreover, we have observed a considerably nonuniform sampling, which is attributed to production variations. However, since the nonuniformity appears systematic, reconstruction methods may still take advantage of the superior sampling rate. The proposed method of interleaved sampling using delay lines may benefit from more advanced technology, and further investigation and refinement may enable sampling of high-frequency signals for coherent reconstruction in integrated sensor systems like radar or LIDAR.

Appendix A

Acronyms

A/D Analog-to-Digital

BBB Beaglebone Black

CMOS Complementary Metal–Oxide–Semiconductor

CTBV Continuous-Time Binary-Value

CV Coefficient of variation

D/A Digital-to-Analog

DAC Digital-to-analog Converter

DFF D flip-flop

DSP Digital Signal Processing

DUT Device under test

ENOB Effective number of bits

EMR Electromagnetic Radiation

GPIB General Purpose Interface Bus

I/O Input/Output

LED Light-emitting diode

LIDAR Light Detection and Ranging

LSB Least significan bit

LVT Low Threshold Voltage (transistors)

MC Monte Carlo analysis

MUX Multiplexer

NIR Near-infrared light

PISO Parallel-in/Serial-out (shift register)

PRF Pulse Repetition Frequency

PWM Pulse Width Modulation

QE Quantization Noise

RADAR Radio Detection and Ranging

RDF Random Dopant Fluctuation

RF Radio Frequency

RMS Root mean square

SCPI General Purpose Interface Bus

SNR Signal-to-noise ratio

SoC System on Chip

SPI Serial Peripheral Interface

TCP/IP a standard transmission protocol

TDDB Time-dependent gate oxide breakdown

ToF Time-of-Flight

TSMC Taiwan Semiconductor Manufacturing Company

Appendix B

Paper (unpublished)

High-speed Sampling in nanometer CMOS

E. Ulvestad, K.G Kjeldgård, T. Moradi Khanshan, D. T. Wisland, T. S. Lande

To be submitted for publication at NorCAS 2018

My contributions: Implementation and measurements of the presented solution as well as figures in paper.

High-speed Sampling in nanometer CMOS

E. Ulvestad, K.G Kjeldgård, T. Moradi Khanshan, D. T. Wisland, T. S. Lande

***Abstract*—In this paper we will present a high-speed sampling system intended for use in a depth selective or depth resolved spectroscopic LIDAR for transcutanuous blood assessment. A continuous-time (CT) sampling solution enabling sampling rates close to 100GHz, close to two orders of magnitude faster than standard clocked systems. Based on experimental verification in 90nm CMOS technology, opportunities and limitations of CT sampling systems is analyzed.**

I. Introduction

Fundamentally we do not have any means of coherent conversion of sensed light signals by sampling electronics due to the high frequency of light. The sophistication and computational power of microelectronics (nanoelectronics these days) have evolved rapidly and by adding photosensors to the silicon die, integrated, on-chip light sensing is feasible as demonstrated by compact and high-quality cameras now available in all sorts of handheld gadgets. Assuming camera-like microelectronics is explored we may implement a spatial 1D or 2D photo detection system. Generating a spectral map of the projected light (image) requires optical elements with wavelength filters. Fortunately new technological solutions exits like linear variable optical filters (LVFs). Assuming on-chip spectral photosensing is feasible, next challenge is to measure the time-of-flight (ToF) of the reflected light. In radar systems a radio beam is transmitted and the ToF of the backscattered signal is measured by sophisticated, high-speed sampling hardware. The idea is to modulate the light with a suitable RF-signal matching the achieved sampling rate (Nyquist limit). Unlike transmission of RF signal, modulated light does not have to comply to strict regulations schemes (FCC, ETSI).

In principle similar solutions may be applied to light using modulated light-pulses reflected (backscattered) from a specific depth inside the body. Seeking existing solutions we find time-of-flight cameras are already available even integrated in CMOS technology [1] [2]. However, available systems are either too large or lack depth resolution required for appropriate selectivity. Approximating the light propagation through the body to be in the order of 10^8m/s, we find 1mm resolution to require in the order of 100GHz sampling rate. Such high sampling frequencies seem to be unreachable in standard clocked CMOS systems as most commercial systems are peaking at a few GHz clock rate. However, at the same time technology is getting better with reduced size and faster digital gates. A typical gate-delay of nanometer technology is better than 10 picoseconds (ps) and getting faster for finer pitch. By substituting the clock synchronization by delay-based timing, systems may operate in continuous time with an operating speed limited only by inherent gate delay of the technology (10ps=100GHz). Handling binary coded signals as 32 or 64 bits words are really hard without the clock, but if we limit the code to just one bit, usable systems are implementable. The name “Continuous-Time Binary-Value (CTBV)” indicates signals coded in continuous time using just a single bit (bitstream) [3]. With CTBV coding we even regain the speed scaling provided by technology downsizing. The CTBV coding-scheme was originally developed for single-chip pulsed radar implementation [4], however, the high-speed sampling and signal-processing solutions are suitable for LIDAR implementation as well.

The ideas of CTBV coding as a “more-than-Moore” technology are published in [3], explaining the fundamentals and versatility of the new design approach. Working with a silicon system, without using clocks (i.e. continuous time), increases processing speed while at the same time saving the power normally required for clocking. For temporal sampling and sequencing inherent gate-delay is used. Added design challenges of production variations and temperature dependencies are handled with calibration and tuning before and during operation [5]. However, process variations is limiting the timing precision of signals propagating through a delayline and add restrictions both on speed and delayline lengths. In the following we have explored parallel delaylines for increased sampling speed of 1-bit sampling systems and experimentally verified equivalent sampling rates towards 100GHz is feasible in 90nm CMOS.

II. High-speed sampler architecture

Constructing a high speed sampling system in the order of 100GHz sampling rate cannot be achieved with standard chip design procedures and in this paper we are exploring the CTBV coding scheme combined with the following requirements:

- Signal transmission is repeatable, preferably with a high repetition rate. In single-chip radar/LIDAR systems this requirement is easy to maintain.
- A coherent signal reconstruction is carried out using swept threshold sampling in combination with CTBV coding.

Combining repetition with single-bit coding and swept-threshold sampling, a complete signal reconstruction is feasible by accumulating the 1-bit samples in accumulators (registers) during the sampling period. In figure 1 the conceptual architecture of a LIDAR is shown. The light is emitted as a pulse. The primary function is to measure Time-of-Flight (ToF) of the light pulse and like radar signalling, suitable modulation of the pulse is used for adaptation to a suitable frequency band set by regulations (FCC,ETSI). However, in a

LIDAR light modulation is less restricted and may be adapted for optimal performance.

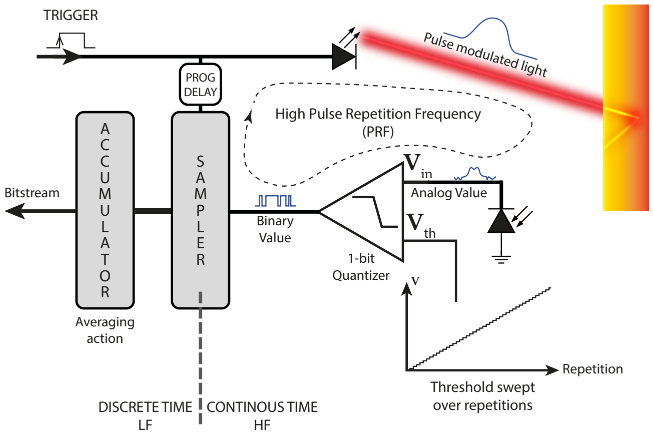


Fig. 1: Overall system architecture

Conceptually we may consider the proposed LIDAR to be a modified radar with a RF-to-light transducer/modulator (laser-diode or LED) on the output and a light-to-RF transducer/demodulator (photo-diode) on the input. In this way we may use radar analysis to determine ToF measurement quality. The down-range resolution of backscattered radar signals from static targets is proportional to the signal bandwidth. Since signal bandwidth is inverse proportional to the signal duration in time domain, a short/fast light pulse will provide the highest depth resolution. Fortunately, wide signal bandwidth are allowed for modulated light. However, bandwidth limitation of transducers as well as CMOS technology is normally limiting signal modulation bandwidth. An additional opportunity is to explore time-varying (moving) radar targets. In medical sensing the pulsating movement of blood vessels due to heart beats would constitute a moving target. Provided a coherent received signal recovery, these movements will occur as phase variations and the detected movement resolution is proportional to the Signal-to-Noise ratio (SNR) of the recovered signal. To some extend increased dynamic target resolution may be traded for reduced signal bandwidth. Typically a realistic RF signal is a modulated RF carrier as indicated in Figure 2. A longer pulse duration will improve phase SNR, but reduce static resolution. An addition, depth resolution in body tissue is improved by $\sqrt{relativepermittivity}$ due to reduced penetration speed of EM waves. Summing up, although we are band-limited by CMOS technology, we may combine static target detection for coarse filtering (time-gating) and explore dynamic features for fine resolution. Since signal repetition is used for pulse reconstruction we also get significant processing gain in the accumulation process improving recovered signal SNR significantly. Sensitivity in the order of millimeters in air is reported [6], [7]. Assuming extremely weak backscattering from subcutaneous reflectors, additional SNR improvements may be achieved by temporal oversampling. SNR is increased with $3dB$ for each doubling of sampling rate.

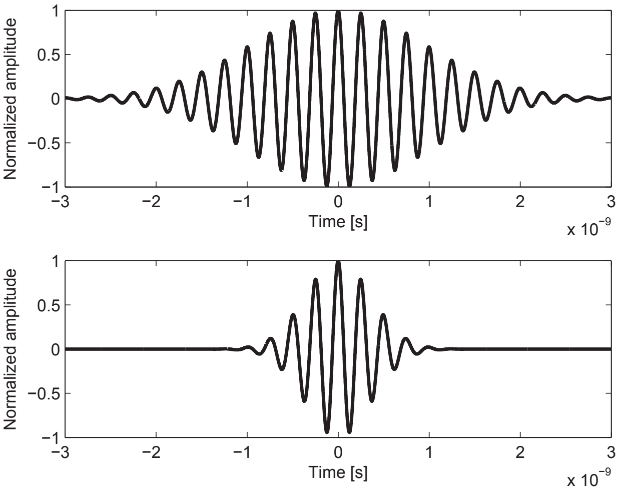


Fig. 2: Balancing phase accuracy with resolution

In this paper we are aiming for a extremely high speed sampling system of a short time window (frame) covering at least the pulse duration. Preferably a longer frame is desirable detecting reflection at different depths.

III. Implementation

Based on the previous assessment we are aiming for highest possible sampling rate of a CTBV-coded bit sequence. The proposed architecture is shown in Figure 3. As indicated the single-bit samplers are simply D-flip-flops. For fast sampling the flip-flops are trigged by a tapped, digital delayline. Each unit delay of the delayline is double inverters enabling latching on the same positive transition of a short pulse propagating through the delayline. Since inverter delay is reduced with finer technology pitch, the sampling speed would benefit from high-end process. However, for cost reasons a 90nm TSMC low-power CMOS process with a nominal inverter delay of $10ps$. Provided a sampling system may explore the speed of an inverter, an equivalent sampling rate in the order of $100GHz$ may be reached. When incorporated in a real implementation, additional input/output loads are reducing speed somewhat. Simulations as well as measurements indicate an accumulated double inverter delay including loads to be in the order of $30ps$.

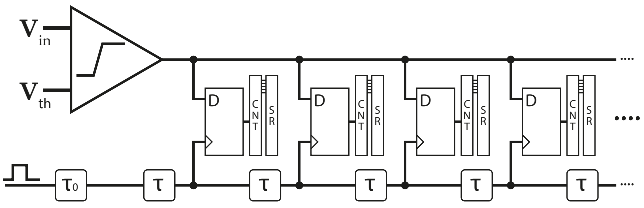


Fig. 3: CT sampling system

As indicated in Figure 4 the idea is to use three parallel and interleaved digital delaylines each with different initial delay enabling triple sampling rate. Achieving this improvements is not possible by design only and the initial delays need to be calibrated. We have explore back-gate or back-biasing of the first inverter as indicated in Figure 5. In addition three different initial delay structure are carefully designed with initial delays of $40ps, 50ps$ and $60ps$ respectively.

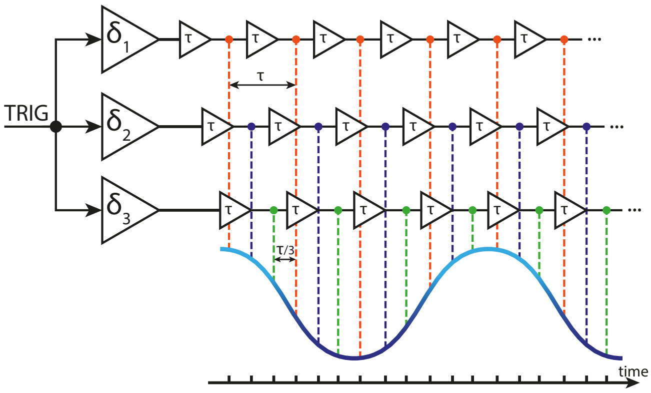


Fig. 4: Interleaving principle

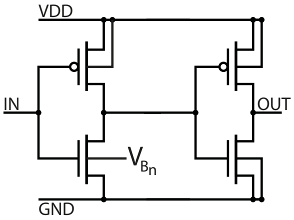


Fig. 5: Back-gate biasing for calibration

An additional challenge is the accumulated unit delay mismatch increasing with the depth of the delayline as indicated in Figure 6. This effect is limiting the depth of interleaved delaylines by reducing the equidistant sampling and even lead to incorrect sampling sequence. In this conceptual work we intend to estimate the achievable depth using delayline interleaving.

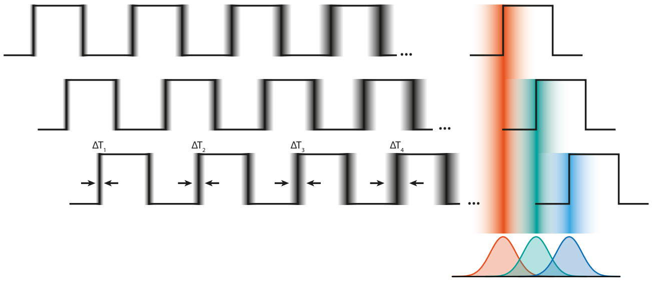


Fig. 6: Accumulated mismatch

IV. Simulations of key elements

In order to reach our design goals extensive simulations including Monte Carlo analysis is required.

A. Unit delay element

The unit delay element used in the digital delayline is a key component and need special design consideration for optimal performance. Target unit delay is $30ps$ with minimal production spread and sharp transitions minimizing jitter. Taking advantage of the positive edge trigged D-FF controlled by the delayline, asymmetric inverter design may be explored as indicated in Figure 7. Another consideration is to maintain sufficient pulse width of the short trigging pulse all through the delay line preventing trigger pulse from disappearing. In very deep delaylines significant design tuning is required [**?**]. However, we are aiming for fairly short delaylines with 16 unit delay elements and may allow for some pulse width variations. Modelling of the unit delay element resulted in the transistor sizes shown in Table I.

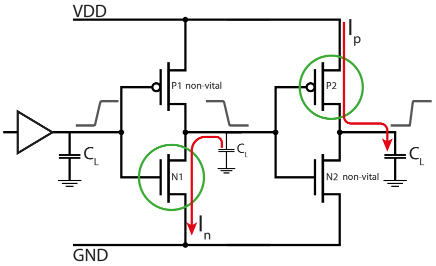


Fig. 7: Asymmetric unit delay design

	Inverter 1		Inverter 2	
	NMOS	*PMOS*	*NMOS*	*PMOS*
fingers n	4	4	4	4
W_n	$950nm$	$1.1\mu m$	$525nm$	$1.22\mu m$
W_{tot}	$3.8\mu m$	$4.4\mu m$	$2.1\mu m$	$4.88\mu m$
L	$100nm$	$100nm$	$100nm$	$100nm$
W/L	38	44	21	~49
$\frac{Wp}{Wn}$	1.16		2.32	

TABLE I: Delay element transistor dimensions

Monte Carlo simulations is giving the results shown in Figure 8.

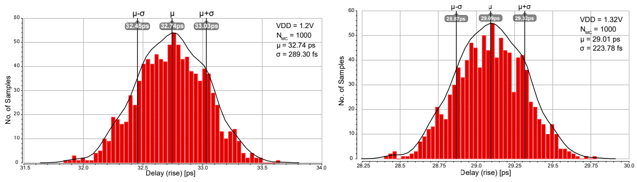


Fig. 8: Monte Carlo simulations of unit delay between raising edges with 1.2V and 1.32V rail voltage

As indicated power supply voltage tuning may be used to calibrate the delayline close to $30ps$ as 1.2V supply voltage is estimated to $32.7ps$ with less that $1ps$ variations and with 1.32V expected unit delay is $29ps$. Also mismatch deviations seem to be acceptable.

V. Measurements

In order to assess the implemented high speed sampler, accurate timing instrumentation is required. We used a RF signal generator (RS SMF 100A) with accurate phase offset adjustments down to $\pm 0.1°$, equal to $\approx 278fs$ at $1GHz$. An RF signal is used as input using an embedded Schmitt-trigger to generate an on-chip sharp transition travelling down the delaylines. By shifting this sharp transition with incremental phase shifts, the sampled results of the accumulators will catch the movement of the transition along the delaylines as indicated in Figure 9

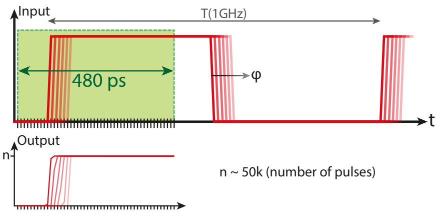


Fig. 9: Delayline measurements using accurate phase shifting of sharp transition

The sampled results are accumulated in 48 registers (16 accumulators for each of the three delaylines). An on-chip SPI interface controlled by a microcontroller is used to read out accumulators for analysis. In order to improve on-chip sampled measurement accuracy, measurements are repeated.

Measured unit delays of sixteen unit delays of each delayline is presented in Figure 10.

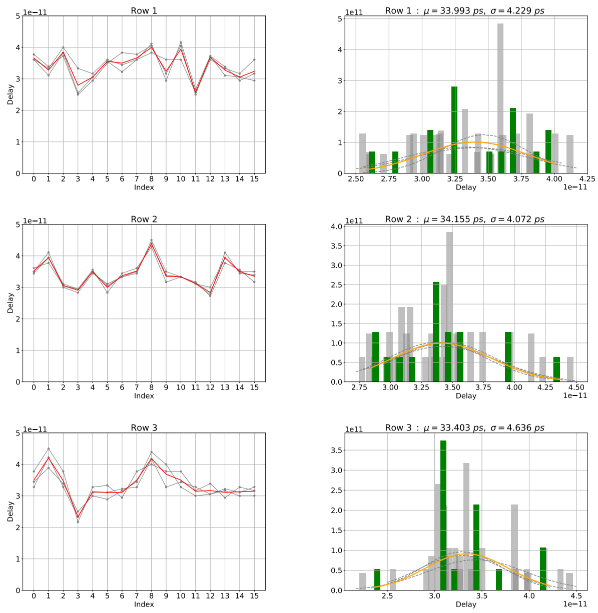


Fig. 10: Measured unit delay of the three delaylines.

For each delayline, unit delays were measured three times. To the left of Figure 10 the red traces are the average of these three measurements while , the green traces are the average in the right-hand traces. As may be seen the average unit delay is $\approx 34ps$, somewhat higher than expected and the spread is more than $4ps$. Apparently additional capacitive loads appear in fabricated chips not captured by post layout simulations. Also, standard deviation is larger than expected possibly indicating even larger devices are required.

Addressing the interleaved delaylines as a temporal sampling system require sequential and ideally equidistant sampling of the incoming signal. Measuring the sequence of an untuned system is shown in Figure 11.

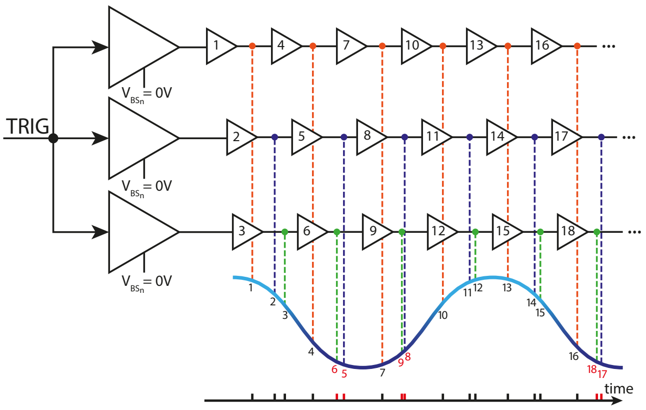


Fig. 11: Interleaved delaylines without tuning.

As may be observed in Figure 11, The three interleaved delaylines give far from equidistant sampling and is even losing sequential order. These irregular sampling errors are primary due to production variations with giving a normal distributed variation in unit delay. Since production variations are static, we may we may use back-gate tuning on the initial delays and establish at least sequential sampling as shown in Figure 12. Although a correct sampling sequence is possible, irregular (nonuniform) sampling of the incoming signal occur. However, advanced signal processing techniques are available for quite good signal reconstruction.

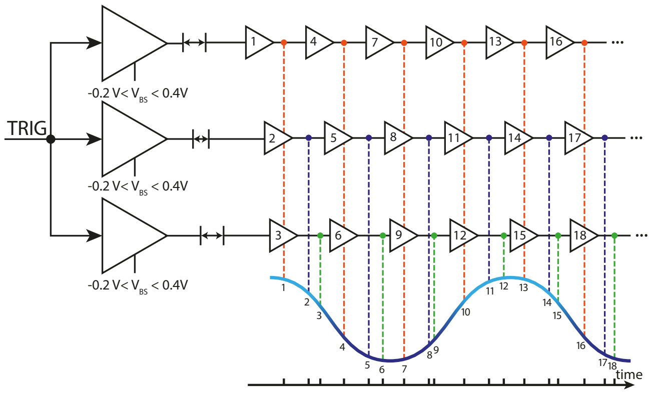

TRIG
$-0.2\,V < V_{BS} < 0.4V$
$-0.2\,V < V_{BS} < 0.4V$
$-0.2\,V < V_{BS} < 0.4V$
1
4
7
10
13
16
2
5
8
11
14
17
3
6
9
12
15
18
time

Fig. 12: Interleaved delaylines with tuning.

Overall system performance is shown in Figure 13 The average unit delay is measured to be $11.9ps$, somewhat higher

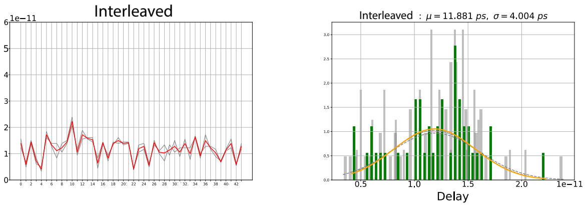


Fig. 13: High speed interleaved delayline measurements.

than targeted and is most likely due to additional parasitic capacitance in real silicon. Also unit delay spread of $4ps$ is larger than expected and is hard to explain. We have carefully examined our instrumentation and evaluated our test-benches for measurement errors and are quite confident. Still a variation from $4ps$ to $23ps$ should be beyond variations in production spread. The complete high speed sampler has a measured, peak power consumption of $< 3.4mW$ indicating the proposed CTBV design approach provide almost two orders of magnitude faster sampling rate than a strictly clocked system even with low power consumption.

VI. Conclusion

In this paper we have presented a low-power and high-speed sampler solution implemented and measured in 90nm TSMC technology. Three interleaved delaylines are combined with tuneable offsets for proper alignment. We were able to measure sequentially correct 48 points with a equivalent sampling rate of $84GHz$. Quite irregular sampling occur due to production variations, but reconstruction methods may still take advantage of the superior sampling rate. The proposed method of interleaved sampling using delaylines may take advantage of more advanced technology and further investigation and refinement may enable sampling of high-frequency (RF) signals for coherent reconstruction in integrated sensor systems like radar or LIDAR.

References

[1] "Pmd technologies home." http://www.pmdtec.com/news_media/news/pico_xs.php, 2008.

[2] "Fotonic home." http://www.fotonic.com/content/Products/Default.aspx.

[3] H. A. Hjortland and T. S. ande, "Ctbv integrated impulse radio design for biomedical applications," *IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS*, vol. 3, pp. 79 – 88, Apr. 2009.

[4] "Novelda home." https://www.xethru.com.

[5] M. Z. Dooghabadi, H. A. Hjortland, and T. S. Lande, "High precision calibrated digital delay element," *Electronics Letters*, vol. 47, pp. 564–565, April 2011.

[6] N. Andersen, K. Granhaug, J. A. Michaelsen, S. Bagga, H. A. Hjortland, M. R. Knutsen, T. S. Lande, and D. T. Wisland, "A 118-mw pulse-based radar soc in 55-nm cmos for non-contact human vital signs detection," *IEEE Journal of Solid-State Circuits*, vol. 52, pp. 3421–3433, Dec 2017.

[7] N. Andersen, K. Granhaug, J. A. Michaelsen, S. Bagga, H. A. Hjortland, M. R. Knutsen, T. S. Lande, and D. T. Wisland, "A 118-mw 23.3-gs/s dual-band 7.3-ghz and 8.7-ghz impulse-based direct rf sampling radar soc in 55-nm cmos," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 138–139, Feb 2017.

Appendix C

Raw data output from phase sweep

Is only readable in PDF format.

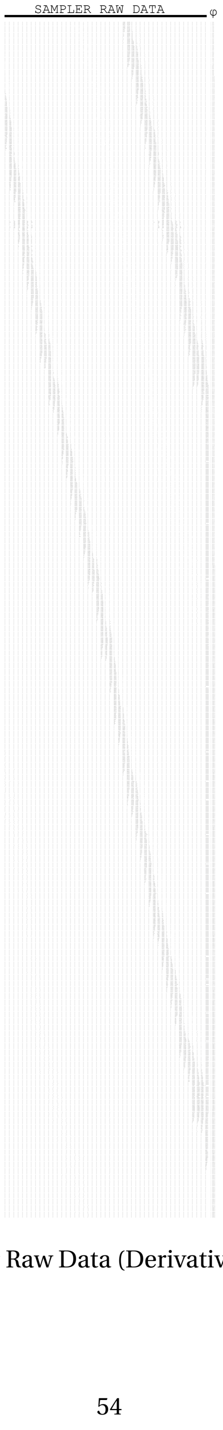


Figure C.1: Sampler Raw Data (Derivative of sampled data).

Appendix D

Layout

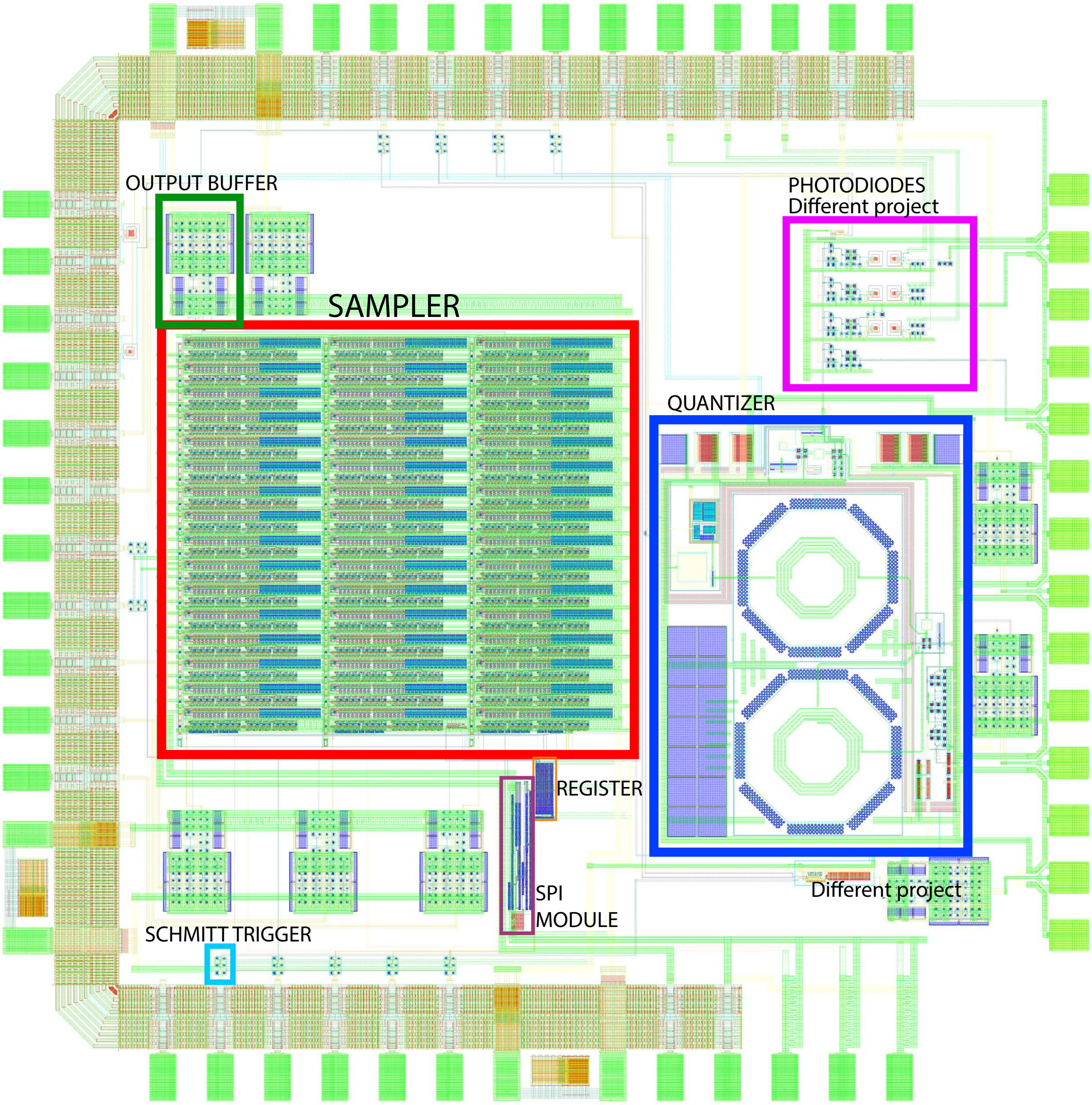


Figure D.1: Padframe

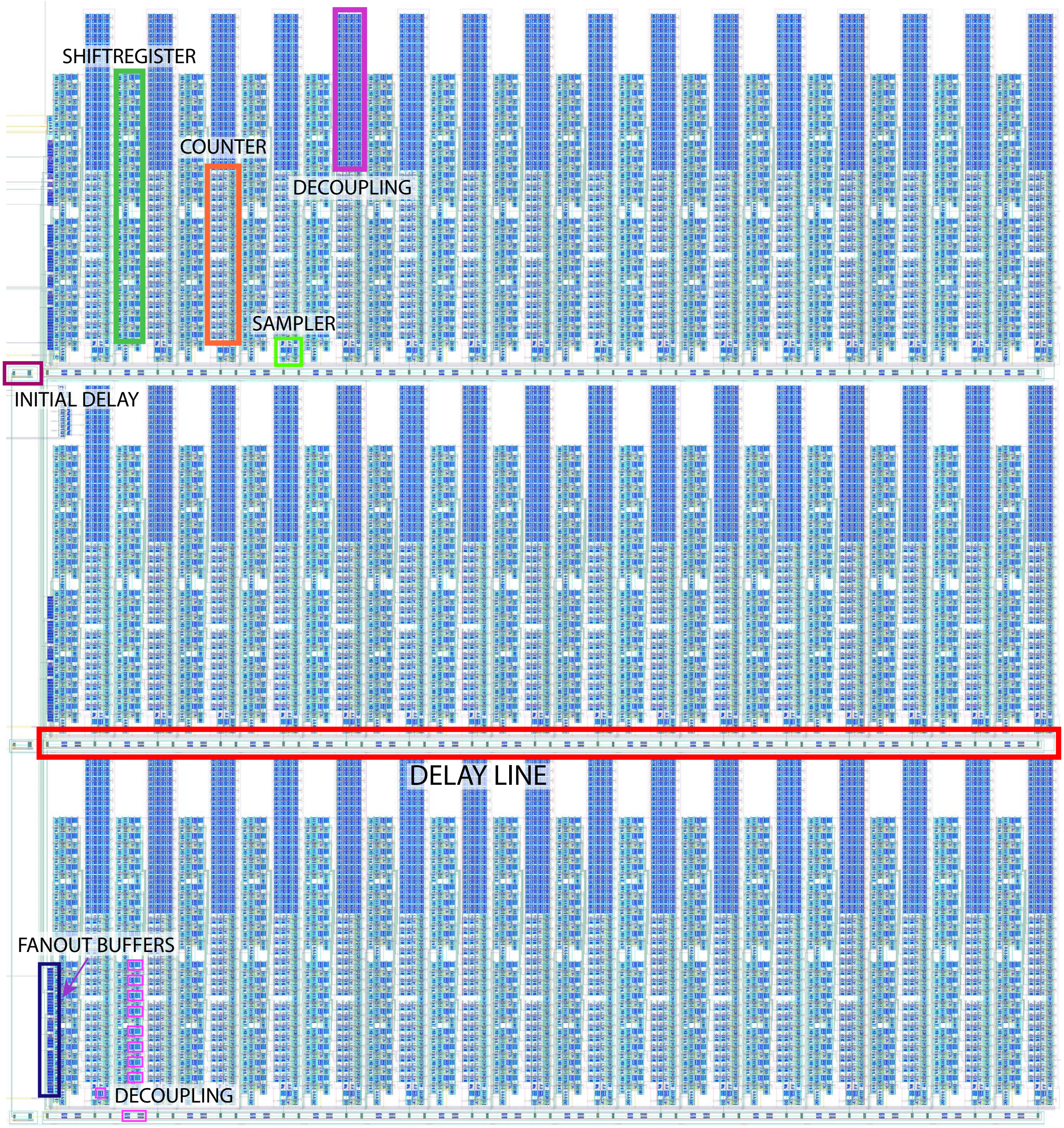


Figure D.2: Sampler Layout

Appendix E

Microscope photographies

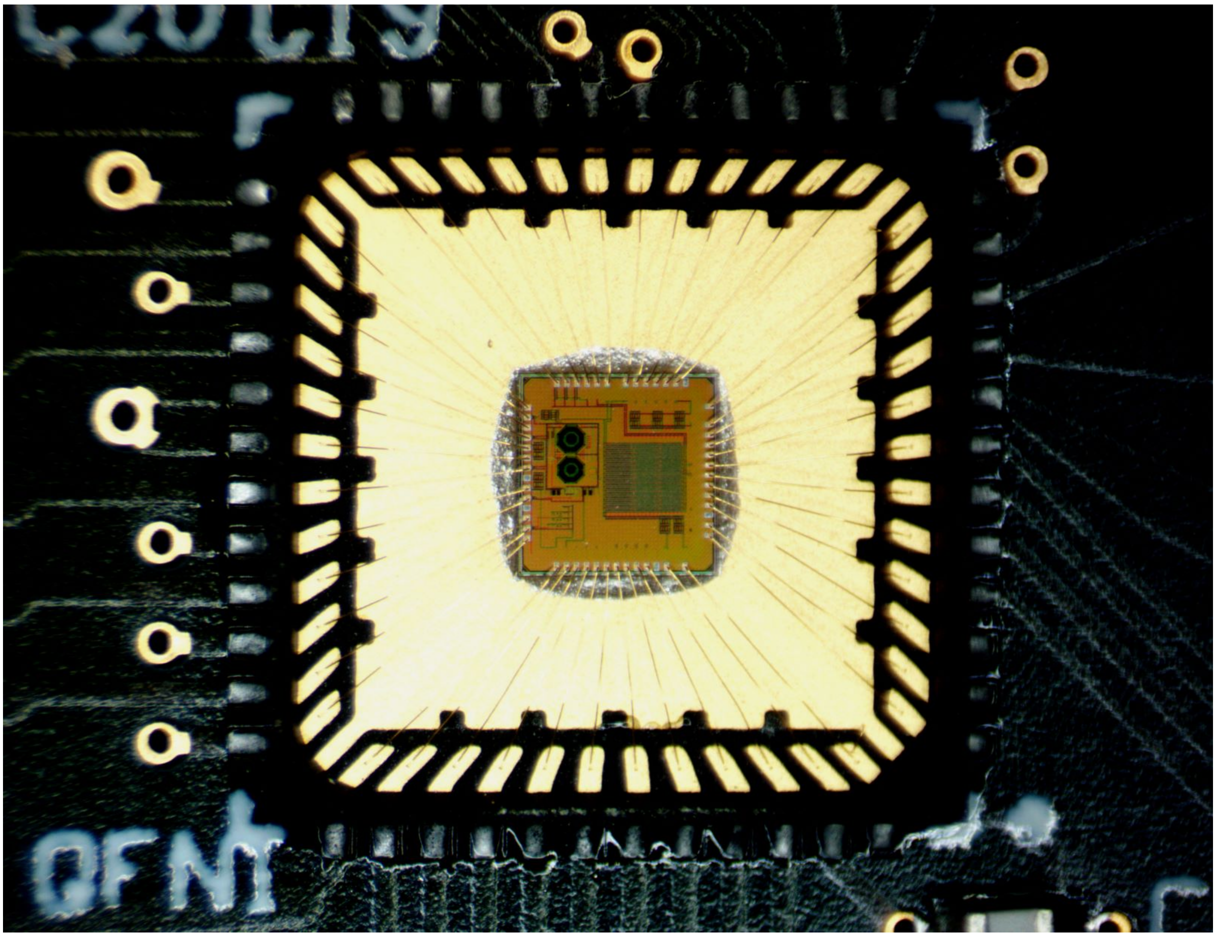


Figure E.1: Package on PCB

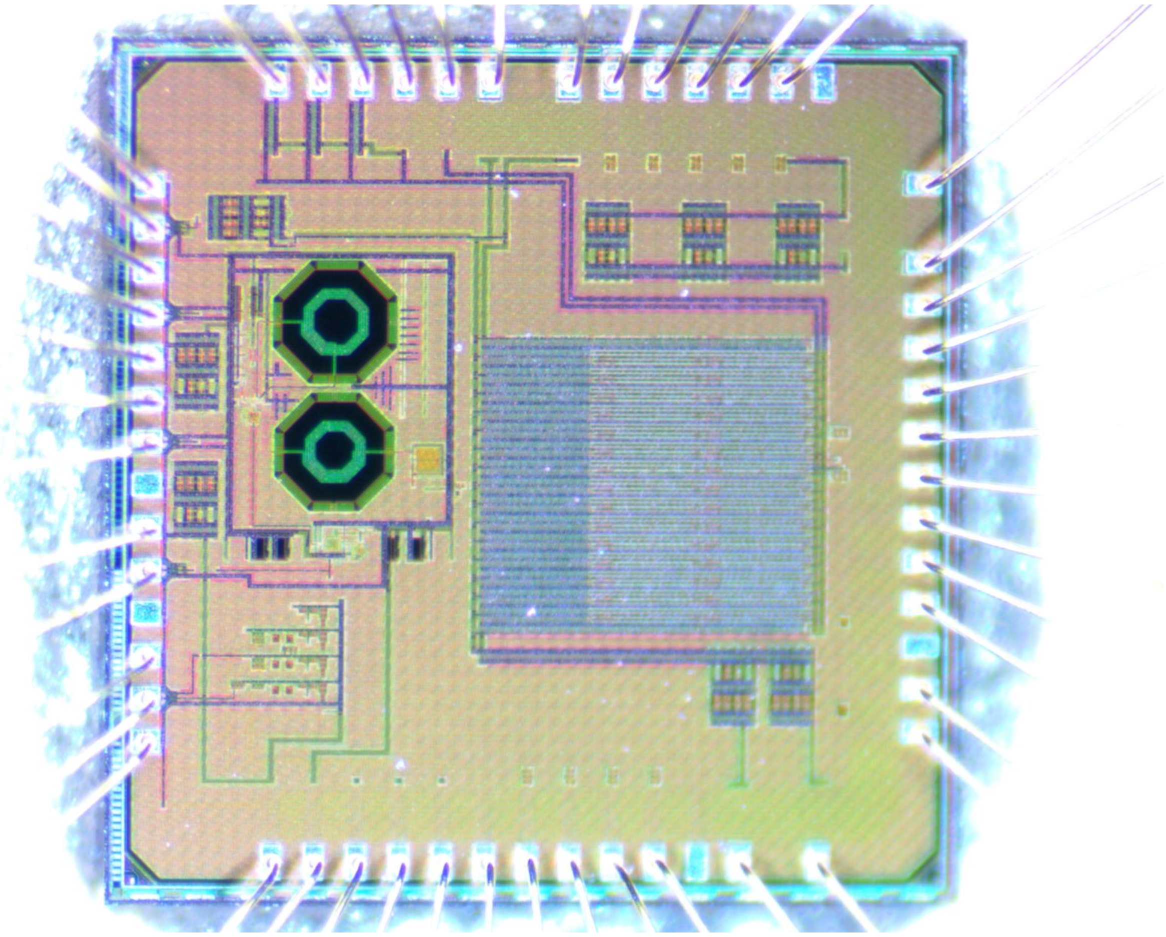
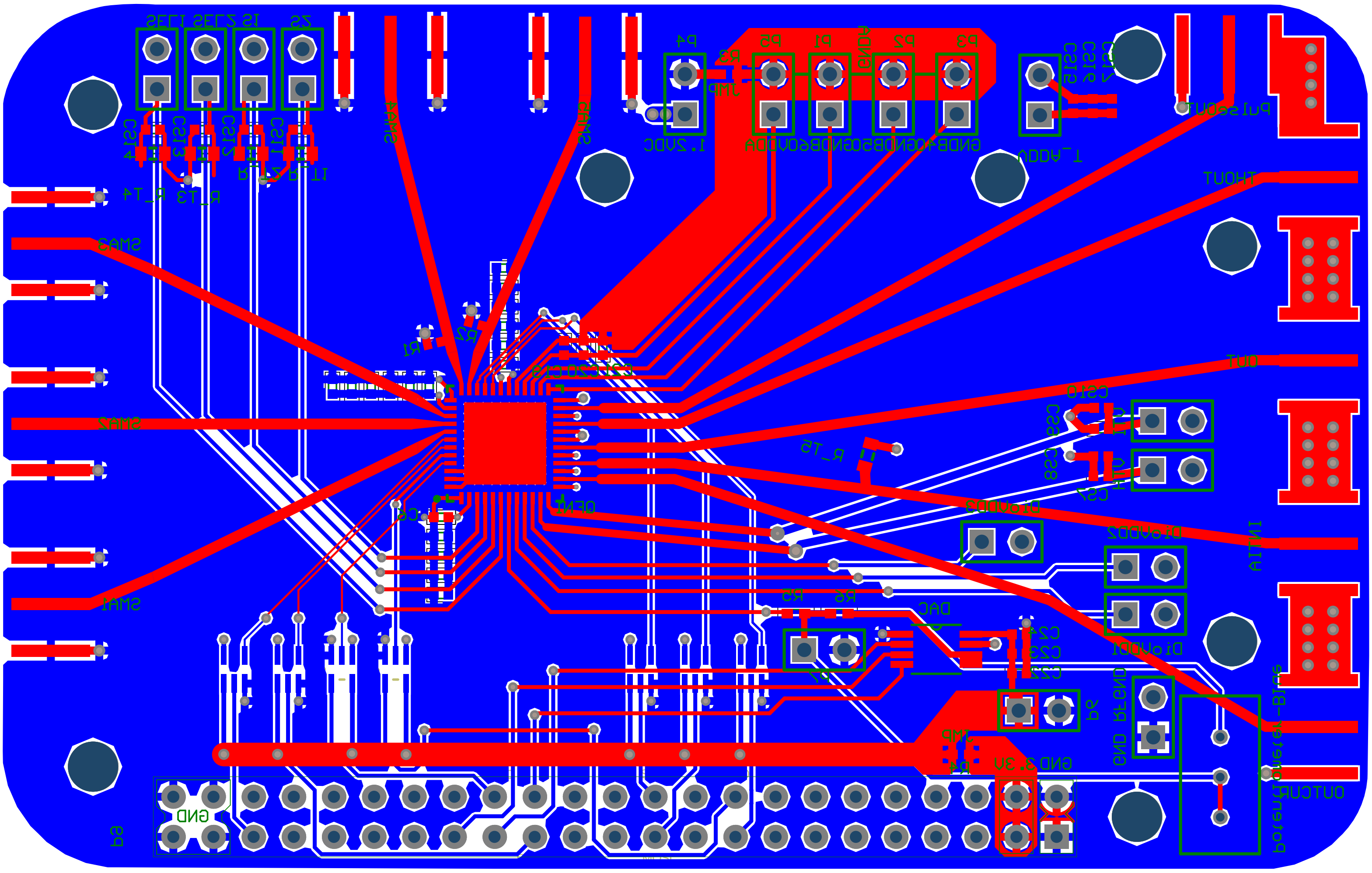


Figure E.2: Successfully implemented chip packaged in QFN48.

Appendix F

PCB

To characterize the sampling system the 4-layer PCB attached below was created. On the board, there is plentiful decoupling, transmission lines with terminations for the RF input signals, voltage translators between chip and Beaglebone Black and the threshold voltage DAC for the quantizer. Additional components are part of the LIDAR project engaged by Tohid Khanshan.


SEL1
SEL2
S1
S2
SMA1
SMA2
SMA3
P1
P2
P3
P4
P5
P6
P9
1.2VDC
R_T1
R_T2
R_T3
R_T4
R_T5
R1
R2
R5
R6
DAC
C22
C23
C24
GND
PulseOUT
INTOUT
OUT
AIN1
OUTCUR

Bibliography

[1] Rebecca Kortum and Eva Sevick. Quantitative Optical Spectroscopy for Tissue Diagnosis. 47:555–606, 02 1996.

[2] N. A. Freebody, A. S. Vaughan, and A. M. Macdonald. On optical depth profiling using confocal raman spectroscopy. *Analytical and Bioanalytical Chemistry*, 396(8), Apr 2010.

[3] Fotonic. *Available at.* `http://fotonic.com`.

[4] PMD Technologies. *Available at.* `http://pmdtec.com`.

[5] Håkon André Hjortland. *Sampled and continuous-time 1-bit signal processing in CMOS for wireless sensor networks.* PhD thesis, Oslo, 2016.

[6] H. A. Hjortland and T. S. B. Lande. CTBV Integrated Impulse Radio Design for Biomedical Applications. *IEEE Transactions on Biomedical Circuits and Systems*, 3(2):79–88, April 2009. ISSN 1932-4545.

[7] Silicon Labs. Improving ADC Resolution by Oversampling and Averaging, 2013.

[8] N. Andersen, K. Granhaug, J. A. Michaelsen, S. Bagga, H. A. Hjortland, M. R. Knutsen, T. S. Lande, and D. T. Wisland. A 118-mW Pulse-Based Radar SoC in 55-nm CMOS for Non-Contact Human Vital Signs Detection. *IEEE Journal of Solid-State Circuits*, 52(12):3421–3433, Dec 2017. ISSN 0018-9200. doi: 10.1109/JSSC.2017.2764051.

[9] D.G. Luchinsky, R. Mannella, P.V.E. Mcclintock, and N.G. Stocks. Stochastic resonance in electrical circuits. i. conventional stochastic resonance. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 46(9):1205–1214, September 1999. ISSN 1057-7130.

[10] Øystein Bjørndal. *Single bit radar systems for digital integration.* PhD thesis, Oslo, 2017.

[11] Novelda. *Available at.* `http://novelda.no`.

[12] R. Jacob Baker. *CMOS Circuit Design, Layout, and Simulation, 3rd Edition.* Wiley, Hoboken, NJ, 2010.

[13] M. Z. Dooghabadi, H. A. Hjortland, and T. S. Lande. High precision calibrated digital delay element. *Electronics Letters*, 47(9):564–565, April 2011. ISSN 0013-5194. doi: 10.1049/el.2011.0395.

[14] O. Dahl, H. A. Hjortland, T. S. Lande, and D. T. Wisland. Close range impulse radio beamformers. In *2009 IEEE International Conference on Ultra-Wideband*, pages 205–209, Sept 2009. doi: 10.1109/ICUWB.2009.5288755.

[15] P. R. Kinget. Device mismatch and tradeoffs in the design of analog circuits. *IEEE Journal of Solid-State Circuits*, 40(6):1212–1224, June 2005. ISSN 0018-9200.

[16] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1439, Oct 1989. ISSN 0018-9200.

[17] M. Quarantelli, S. Saxena, N. Dragone, J. A. Babcock, C. Hess, S. Minehane, S. Winters, Jianjun Chen, H. Karbasi, and C. Guardiani. Characterization and modeling of MOSFET mismatch of a deep submicron technology. In *International Conference on Microelectronic Test Structures, 2003.*, pages 238–243, March 2003.

[18] T. Mizuno, J. Okumtura, and A. Toriumi. Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's. *IEEE Transactions on Electron Devices*, 41(11):2216–2221, Nov 1994. ISSN 0018-9383.

[19] Martin Wirnshofer. *Variation-Aware Adaptive Voltage Scaling for Digital CMOS Circuits*, volume 41 of *Springer series in advanced microelectronics.* 2013. ISBN 1-299-40839-7.

[20] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen. Modeling statistical dopant fluctuations in MOS transistors. *on Electron Devices*, 45(9):1960–1971, Sep 1998. ISSN 0018-9383. doi: 10.1109/16.711362.

[21] P. G. Drennan and C. C. McAndrew. Understanding MOSFET mismatch for analog design. *IEEE Journal of Solid-State Circuits*, 38(3):450–456, March 2003. ISSN 0018-9200.

[22] T. Sakurai and A. R. Newton. Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, Apr 1990. ISSN 0018-9200.

[23] Texas Instruments. CMOS Power Consumption and Cpd Calculation, 1997.

[24] A. Strak and H. Tenhunen. Investigation of Timing Jitter in NAND and NOR Gates Induced by Power-Supply Noise. In *2006 13th IEEE International Conference on Electronics, Circuits and Systems*, pages 1160–1163, Dec 2006.

[25] Bilal Abdulrazzaq, Izhal Abdul Halin, Shoji Kawahito, Roslina Sidek, Suhaidi Shafie, and Nurul Yunus. A review on high-resolution CMOS delay lines: towards sub-picosecond jitter performance. *SpringerPlus*, 5 (1):1–32, 2016. ISSN 2193-1801.

[26] Shanthi Sudalaiyandi. *Continuous-time symbol detection for high-precision RF localization suitable for wireless health care.* PhD thesis, Oslo, 2014.

[27] T. Mizuno. Influence of statistical spatial-nonuniformity of dopant atoms on threshold voltage in a system of many MOSFETs. *Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers*, 35(2):842–848, February 1996. ISSN 00214922.

[28] K. von Arnim, E. Borinski, P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, and C. Pacha. Efficiency of Body Biasing in 90-nm CMOS for Low-Power Digital Circuits. *IEEE Transactions of Solid-state Circuits*, 40(7):175–178, July 2005.

[29] Behzad Razavi. Design of analog CMOS integrated circuits, 2017.

[30] T. Charania, A. Opal, and M. Sachdev. Analysis and Design of On-Chip Decoupling Capacitors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 21(4):648–658, April 2013. ISSN 1063-8210.

[31] Beaglebone Black. *Available at.* `https://beagleboard.org/black`.

[32] Farokh Marvasti. Nonuniform sampling : Theory and practice, 2001.